REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY) 31-01-2017	2. REPORT TYPE Final Report	3. DATES COVERED (From - To) 23-Sep-2010 - 31-Oct-2016

4. TITLE AND SUBTITLE
Final Report: Continuation Study: A Systems Approach to Understanding Post-Traumatic Stress Disorder

5a. CONTRACT NUMBER
W911NF-10-2-0111

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
365000

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHORS
Francis J. Doyle III, Kai Wang, Linda Petzold

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Santa Barbara
3227 Cheadle Hall
3rd floor, MC 2050
Santa Barbara, CA 93106 -2050

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
58488-LS.21

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Post-Traumatic Stress Disorder (PTSD) is a complex anxiety disorder affecting many combat-exposed soldiers. Current diagnosis of PTSD is survey-based and is not used to diagnose stages of the disorder, reliably inform effective treatment strategies, or predict recovery/symptom changes. Thus, there is a need to identify robust biomarkers for accurate diagnosis, prognosis, and evaluation of therapeutics. Using the currently available blood data from the Systems Biology of PTSD Consortium, we sought to provide greater insights into the complex underlying biophysical networks of PTSD using a variety of statistical, machine learning, and dynamic modeling

15. SUBJECT TERMS
Post Traumatic Stress Disorder, HPA-Circadian-metabolic pathway, methylation, biomarkers

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Francis Doyle
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU	UU		19b. TELEPHONE NUMBER 805-893-8133

Report Title

Final Report: Continuation Study: A Systems Approach to Understanding Post-Traumatic Stress Disorder

ABSTRACT

Post-Traumatic Stress Disorder (PTSD) is a complex anxiety disorder affecting many combat-exposed soldiers. Current diagnosis of PTSD is survey-based and is not used to diagnose stages of the disorder, reliably inform effective treatment strategies, or predict recovery/symptom changes. Thus, there is a need to identify robust biomarkers for accurate diagnosis, prognosis, and evaluation of therapeutics. Using the currently available blood data from the Systems Biology of PTSD Consortium, we sought to provide greater insights into the complex underlying biophysical networks of PTSD using a variety of statistical, machine learning, and dynamic modeling techniques. Primarily, our analysis was completed on an age and ethnicity-matched male cohort of 83 PTSD and 83 combat-exposed control subjects, a preliminary validation cohort for some data types, and a small of cohort of recalled subjects from the original 83-83 cohort. Using this available data, we focused our efforts on five aims: (1) characterization of disease signals, and affected biological pathways in PTSD, (2) development and application of single ‘omic biomarker identification tools, (3) integration of multi-omics datasets for biomarker identification, (4) characterization of DNA methylation-based subtypes of PTSD, (5) development of an HPA-circadianmetabolic
dynamic model, and (6) development of data analysis pipelines for large molecular datasets.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
01/31/2017	20	K. Sriram, Maria Rodriguez-Fernandez, Francis J. Doyle, Attila Gursoy. A Detailed Modular Analysis of Heat-Shock Protein Dynamics under Acute and Chronic Stress and Its Implication in Anxiety Disorders, PLoS ONE, ( ): . doi:
08/31/2016	18	Gunjan Thakur, Bernie Daigle Jr., Meng Qian, Kelsey Dean, Yuanyang Zhang, Ruoting Yang, Taek-Kyun Kim, Xiaogang Wu, Meng Li, Inyoul Lee, Linda Petzold, Francis Doyle III. A Multi-Metric Evaluation of Stratified Random Sampling for Classification: A Case Study, IEEE LIFE SCIENCES LETTERS, ( ): 1. doi:
08/31/2016	19	Vikas Ghai, Kai Wang. Recent progress towards the use of circulating microRNAs as clinical biomarkers, Archives of Toxicology, ( ): . doi:
TOTAL:	**3**	

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received	Paper
TOTAL:	

Number of Papers published in non peer-reviewed journals:

(c) Presentations

- Ruoting Yang, Bernie J Daigle Jr, Linda R Petzold, Francis J Doyle III. Core module network construction for breast cancer metastasis. Presented at 10th World Congress on Intelligent Control and Automation (WCICA). Beijing, China. July 6-8 2012

- Gunjan S Thakur, Bernie J Daigle Jr, Linda R Petzold, Francis J Doyle III. A multivariate ensemble approach for identification of biomarkers: application to breast cancer. Presented at 19th World Congress of the International Federation of Automatic Control (IFAC). Cape Town, South Africa. August 24-29 2014.

- Kai Wang. Characterization of cell-free RNA in circulation Presented at Molecular Triconference. San Francisco, CA. March 16, 2016.

- Kai Wang. The need of standardization on measuring circulating RNA. Presented at Molecular Tri-conference. San Francisco, March 16, 2016.

- Yong Zhou. Organ-specific proteins in biomarker discovery. Presented at Biomarker Summit. San Diego, CA. March 21-23, 2016.

- Min Young Lee. Discovery of integrative biomarkers for Post-traumatic stress disorder with brain imaging, endocrine, and proteomic features. Presented at the Cascadia Proteomics Symposium. Seattle, WA. July 11-12, 2016.

Number of Presentations: 6.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper
TOTAL:	

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received		Paper
08/31/2011	3.00	Ruoting Yang, K. Sriram, Francis J. Doyle III. Control Circuitry for Fear Conditioning Associated with Post-Traumatic Stress Disorder (PTSD), 49th IEEE Conference on Decision and Control. 15-DEC-10, . : ,
TOTAL:		**1**

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
08/31/2011	4.00	John K. House, Robert L. Sinsheimer, William R. Shimp, Michael J. Mahan, Douglas M. Heithoff. Intraspecies variation in emerging hyperinfectious bacterial strains in nature, PLoS Pathogens (08 2011)
TOTAL:		**1**

Number of Manuscripts:

Books

Received	Book
TOTAL:	

Received	Book Chapter
TOTAL:	

Patents Submitted

None

Patents Awarded

None

Awards

None

Graduate Students

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Francis J. Doyle III	0.17	
FTE Equivalent:	**0.17**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME
Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED
Inyou Lee	0.35
Li Tang	0.75
Shizhen Qin	0.60
Yong Zhou	0.75
David Baxter	0.20
New Entry	0.00
FTE Equivalent:	**2.65**
Total Number:	**6**

Sub Contractors (DD882)

1 a. The Institute for Systems Biology

1 b. 1441 North 34th Street

Seattle WA 981038904

Sub Contractor Numbers (c): KK1127

Patent Clause Number (d-1): NONE

Patent Date (d-2):

Work Description (e): To identify molecular expression changes occurred during PTSD that can be used in deve

Sub Contract Award Date (f-1): 9/23/10 12:00AM

Sub Contract Est Completion Date(f-2): 10/31/16 12:00AM

---

1 a. The Institute for Systems Biology

1 b. 1441 North 34th Street

Seattle WA 981038904

Sub Contractor Numbers (c): KK1127

Patent Clause Number (d-1): NONE

Patent Date (d-2):

Work Description (e): To identify molecular expression changes occurred during PTSD that can be used in deve

Sub Contract Award Date (f-1): 9/23/10 12:00AM

Sub Contract Est Completion Date(f-2): 10/31/16 12:00AM

---

1 a. Harvard University

1 b. Office for Sponsored Programs
1033 Massachusetts Ave 5th Floor
Cambridge MA 021385369

Sub Contractor Numbers (c): KK1620

Patent Clause Number (d-1): NONE

Patent Date (d-2):

Work Description (e): The Harvard team will lead the informatics analysis and computational modeling efforts.

Sub Contract Award Date (f-1): 7/1/15 12:00AM

Sub Contract Est Completion Date(f-2): 10/31/16 12:00AM

---

1 a. Harvard University

1 b. Holyoke Center, 7th Floor
1350 Massachusetts Avenue
Cambridge MA 021383846

Sub Contractor Numbers (c): KK1620

Patent Clause Number (d-1): NONE

Patent Date (d-2):

Work Description (e): The Harvard team will lead the informatics analysis and computational modeling efforts.

Sub Contract Award Date (f-1): 7/1/15 12:00AM

Sub Contract Est Completion Date(f-2): 10/31/16 12:00AM

Inventions (DD882)

Scientific Progress

See Attachment

Technology Transfer

None

Abstract

Post-Traumatic Stress Disorder (PTSD) is a complex anxiety disorder affecting many combat-exposed soldiers. Current diagnosis of PTSD is survey-based and is not used to diagnose stages of the disorder, reliably inform effective treatment strategies, or predict recovery/symptom changes. Thus, there is a need to identify robust biomarkers for accurate diagnosis, prognosis, and evaluation of therapeutics. Using the currently available blood data from the Systems Biology of PTSD Consortium, we sought to provide greater insights into the complex underlying biophysical networks of PTSD using a variety of statistical, machine learning, and dynamic modeling techniques. Primarily, our analysis was completed on an age and ethnicity-matched male cohort of 83 PTSD and 83 combat-exposed control subjects, a preliminary validation cohort for some data types, and a small of cohort of recalled subjects from the original 83-83 cohort. Using this available data, we focused our efforts on five aims: (1) characterization of disease signals, and affected biological pathways in PTSD, (2) development and application of single 'omic biomarker identification tools, (3) integration of multi-omics datasets for biomarker identification, (4) characterization of DNA methylation-based subtypes of PTSD, (5) development of an HPA-circadian-metabolic dynamic model, and (6) development of data analysis pipelines for large molecular datasets.

Table of Contents

1. Background

Post-Traumatic Stress Disorder (PTSD) is a psychological condition occurring in some people after experiencing traumatic events [1-2]. The diagnosis and treatment of PTSD poses a challenging problem for contemporary medicine in that its primary effects are mediated through the brain, and we have a poor understanding of the molecular process underlying the progression of the disease. The assessment methods for PTSD are largely based on clinical interviews and sensory tests; these may be highly subjective tests with significant uncertainty. Therefore, there is an urgent need to develop clinically relevant molecular biomarkers for PTSD diagnosis. The Systems Biology of PTSD Consortium (PIs: Charles Marmar and Marti Jett) has ascertained blood samples from approximately 350 total PTSD and control subjects for molecular data analysis, including genetics, epigenetics, proteomics, metabolomics, transcriptomics, miRNA, and endocrine marker studies. Using a systems biology approach, data across these length scales can be integrated to characterize and refine our understanding of PTSD (**Figure 1.1**) [3]. Primarily, the goals of this consortium are to identify a robust and accurate PTSD biomarker panel from this multi-omic dataset. A summary of the currently available molecular and clinical data is shown in Table 1.1. The subjects have been divided into an initial male discovery cohort of 83 PTSD and 83 age and ethnicity-matched controls, a male validation cohort, a female cohort, and a cohort of subjects returning for a follow-up study (the recall cohort). Some analysis of the 83-83 cohort was divided into training on a 52-52 subset, with validation on the remaining 31-31 subjects.

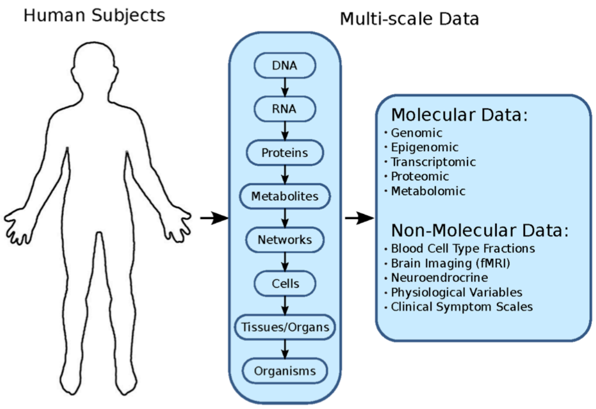


Figure 1.1 Systems biology approach for the study of PTSD: a multitude of data types can be experimentally collected by probing biological systems at different scales. This wealth of data is then further processed by employing mathematical and computational techniques (*e.g.*, statistics, machine learning, computational modeling) to gain insight into the system under study (from [3]).

Table 1.1 Summary of currently available data, as of January 1, 2017.

	# of features	Male Discovery		Male Validating		Female		Recall	
		# PTSD	# Control	# PTSD	# Control	# PTSD	# Control	# PTSD	# Control
Clinical	151	83	83	33	44	22	30	35	33
Clinical Lab	44	81	82	19	26	4	7	29	25
Endocrine	35	82	83	13	20	15	20	29	25
Metabolite	224	82	82	-	-	-	-	-	-
Protein	96	82	80	13	20	19	20	-	-
miRNA	43	71	74	-	-	19	21	-	-
DNA Methylation	429248	81	82	15	22	6	4	13	8
mRNA	50599	76	80	-	-	-	-	-	-
SNP	557423	56	67	1	5	6	11	N/A	N/A

We believe that the blood is a rich resource for disease biomarkers in that it is the lifeline and one of the major communication channels for all cells in the body. In addition, it is easily accessible. A number of blood protein-based disease markers have been discovered and are used in clinics. For example, PSA levels in blood have been routinely used for prostate cancer detection [4] and blood troponin levels for heart diseases [5].

Besides proteins and peptides, RNA, especially microRNA (miRNA), a group of small regulatory RNAs, have recently been detected in circulation [6-9]. A portion of these miRNAs is packaged in lipid vesicles such as exosomes released by cells. Exosomes are vesicles that originate from multi-vesicle bodies in the cells and contain different macromolecules from the originating cell. Recent studies have demonstrated that exosomes in circulation can interact with or be taken up by cells where they produce functional effects. These findings suggest that exosomes may participate in the cell-cell communication network [10-13]. Therefore, disease conditions may affect the spectrum of blood proteins, miRNAs, and exosomal content. We hypothesized that PTSD will affect the spectrum of protein and RNA in plasma and exosomes. Using global profiling technologies, we will be able to identify PTSD-associated proteins and RNAs in circulation. These molecules can be used as biomarkers to detect and stratify patients with PTSD.

In addition to peptides and miRNA, genetic and epigenetic signals may contain signals of risk or resilience. Previous evidence suggests genetic risk of PTSD exists, though the specific links and connections to other comorbid disorders are still unclear [14]. Recently, a few single nucleotide polymorphisms (SNPs) were identified to be significantly associated with PTSD risk [15-16], and may provide insights into functional changes occurring during PTSD development and progression. Beyond genetics, epigenetic changes, specifically DNA methylation, may indicate signals of acute or chronic stress [17]. These epigenetic changes can occur during and after trauma exposure, resulting in altered regulation of gene expression.

Larger macromolecules may also play a role in identifying molecular and cellular signals of PTSD. Genetic and epigenetic changes may propagate through transcription and translation via altered regulation, resulting in changes in larger molecular components, including metabolites, cytokines, and endocrine signals. Increased cellular aging, based on measured telomere length, has been reported in combat veterans with PTSD [18-19]. Other metabolism changes, including mitochondrial dysfunction and changes in oxidative stress have been reported in PTSD and other neuropsychiatric disorders including schizophrenia and depression [20-21]. By integrating data from genetics to more complex macromolecules, greater insight can be gained connecting risk, resilience and progression of PTSD.

The proteogenomic core at ISB and the bioinformatics and modeling core at Harvard are part of the multi-omics effort for the overall Systems Biology of PTSD Consortium with a goal to understand the disease, and to identify and validate PTSD diagnostic markers. Working with Dr. Marmar at NYU Langone Medical Center and Dr. Jett at the U.S. Army Center for Environmental Health Research, ISB's proteogenomic core conducted comprehensive analyses on the plasma proteins and RNAs and the corresponding exosomal RNA from samples obtained from male OIF/OEF (Operation Iraqi Freedom/Operation Enduring Freedom) veterans with or without PTSD. We identified a set panel of proteins and miRNAs blood biomarker candidates that showed good performance to diagnose PTSD. With the finding, we propose to further validate the panel with different cohorts of patients and expand the initial male study to refine the PTSD blood diagnostic panel already identified.

The bioinformatics and modeling core at Harvard develops and applies machine learning, statistics, and dynamic modeling to integrate the multi-omic PTSD datasets in order to: (1) identify robust biomarker panels for PTSD diagnosis from single 'omic and multi-omic datasets, (2) characterize and model the disease mechanism leading to observable phenotypes, (3) explain heterogeneity in PTSD subjects based on novel disease subtypes, stages of progression, or genetic risk groups, and (4) predict potential therapeutic targets.

2. Analysis of emerging disease signals and characterization of PTSD biology

2.1 Genome wide association study analysis

Population-Specific and Trans-Ancestry Variants Associated with PTSD Severity

The influence of genetic factors on PTSD resilience and prognosis has long been recognized. Efforts to map the genetic architecture of PTSD primarily include two major approaches. Early on, candidate-gene studies investigated relationships between disease phenotypes and polymorphisms on targeted genes of interest chosen based on some biological hypothesis. Later, with the advent of the genomic era and availability of SNP arrays, a few genome-wide association studies (GWAS) were carried out to identify genetic variants that are associated with PTSD onset and severity in a hypothesis-neutral manner.

However, perhaps alarmingly, findings from candidate-gene studies are overall disparate from significant GWAS hits. This lack of overlap has been attributed to various technical challenges. On one hand, the biology of PTSD is too little understood to guide a successful and reliable hypothesis-driven approach (candidate gene studies); on the other hand, hypothesis-neutral (data-driven) approaches are still underpowered because they require a large sample size to have a reliably generalizable result.

Here we investigate variants on genes and intergenic loci implicated in previous GWAS and candidate-gene studies. Our dataset is carefully chosen to involve individuals with extreme PTSD severity score [22]. This has primarily three advantages. First, it minimizes inadvertent misdiagnosis. Second, it enhances statistical power. Third, the biological distinction is likely to be more pronounced between the extreme cases.

Genes and intergenic loci from previous PTSD genetic studies

We collected variants located on genes of interest from UCSC Genome Browser (GRCh37/hg19 Assembly) (https://genome.ucsc.edu/). These genes include those previously studied on candidate gene studies [23-24]: ADCYAP1, ADCYAP1R1, ANK3, ANKK1, APOE, BDNF, CAT, CHRNA5, CNR1, COMT, CRHR1, DBH, DRD2, DRD4, DTNBP1, FKBP5, GABRA2, HTR2A, KPNA3, MAOB, NOS1AP, NPY, NR3C1, PRKCA, RGS2, SLC18A2, SLC6A3, SLC6A4, SRD5A2, STMN1, TPH2, and WWC1. Additional genes implicated in previous GWAS studies include: PRTFDC1 (rs6482463) [25], lincRNA AC068718.1 (rs10170218) [26], RORA (rs8042149) [27], TLL1 (rs406001) [28], ANKRD55 (rs159572) and ZNF626 (rs11085374) [29].

Dataset and Statistical Analysis

Our genotype data consists of 147 samples [22]. The initial Illumina OmniExpress Beadchip measures 730,493 SNPs. 557,423 SNPs, including 918 SNPs on the genes of interest, on 147 samples survived the quality-control steps (minimum threshold for minor allele frequency of 0.01, maximum SNP missingness rate of 0.05, maximum individual missingness rate of 0.05, and Hardy-Weinberg equilibrium p-value of 1e-5). Linear regression is performed on current CAPS as the response variable and minor-allele frequency, sex, and three principal components as explanatory variables. This analysis is done on PLINK [30].

We found a nominal significant ($p<0.01$) association between eight variants on six genes (**Table 2.1**). Interestingly, two SNPs on a gene previously implicated in a GWA study, TLL1, and two intronic SNPs on BDNF locus passed nominal significant threshold (**Figure 2.1**).

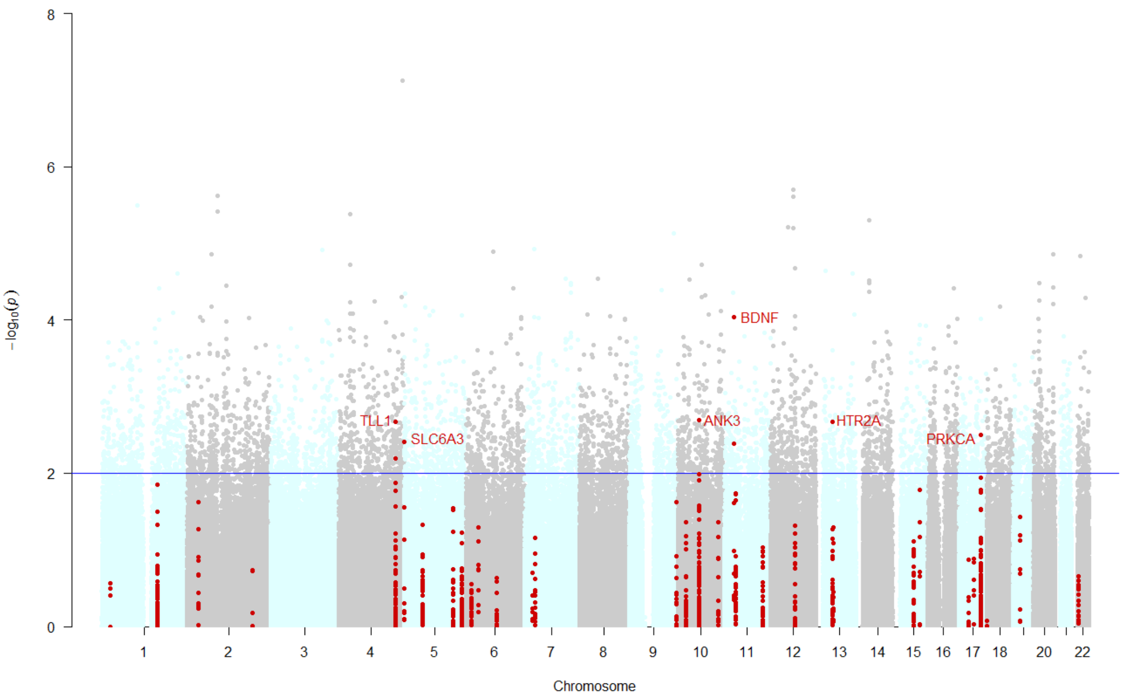


Figure 2.1 Manhattan plot of association between PTSD severity and variants on autosomal chromosomes. SNPs on selected genes are shown in red. The horizontal blue line represents a nominal significance level (p=0.01).

Table 2.1 SNPs with statistically significant association (p<0.01) with PTSD severity.

CHR	SNP	Gene	Function	BP	BETA	p-value
4	rs7696087	TLL1	intron variant	1.67E+08	14.85	0.002115
4	rs4691229	TLL1	intron variant	1.67E+08	12.09	0.006409
5	rs1042098	SLC6A3	3 prime UTR variant	1394815	9.928	0.003887
10	rs10821659	ANK3	intron variant	61793424	10.78	0.001999
11	rs12291063	BDNF	intron variant	27694101	16.74	9.08E-05
11	rs7124442	BDNF	intron variant	27677041	-10.43	0.004098
13	rs1923885	HTR2A	intron variant	47423086	-10.9	0.002149
17	rs17771145	PRKCA	intron variant	64453228	-15.22	0.003171

We further investigated the relationship between genotype of the top SNP (rs12291063) and PTSD severity. We performed the analysis for the three major ancestry groups separately (Hispanics, non-Hispanic Whites and non-Hispanic Blacks) and all samples together (**Figure 2.2**). Our preliminary analysis suggests the genetic architecture of PTSD has population specific components, reflective of the very heterogeneous nature of the disorder.

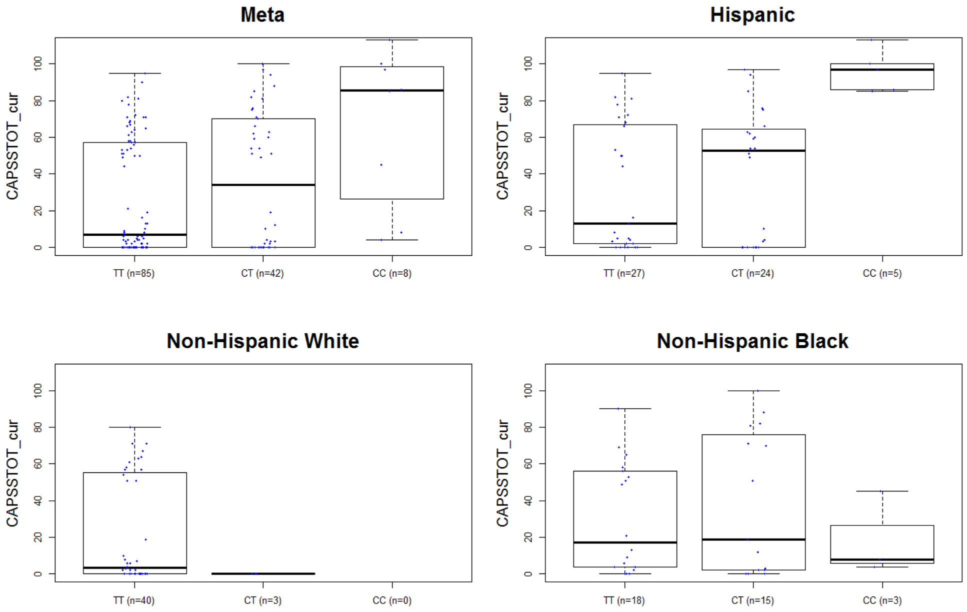


Figure 2.2 Boxplot showing relationship between an intronic SNP of BDNF (rs12291063) and PTSD severity, based on current CAPS total. Combined analysis of all ancestries showed a nominally significant association, p<0.01 (top right). Individual ancestry group analyses showed a nominally significant association only in the Hispanic group.

An appealing functional SNP on BDNF, rs6265 (Val66Met), is widely studied in several psychiatric disorders. On a finding on 42/419 case/control sample from the US Army Special Operations soldiers deployed to OEF/OIF, a recent study reported probable-PTSD is associated with a BDNF functional polymorphism rs6265 (Val66Met) and higher plasma concentration [31]. In our data we found a statistically significant relationship between BDNF blood concentration (**Figure 2.3**) and PTSD severity, but did not find a significant association between rs6265 genotypes and PTSD diagnosis nor severity.

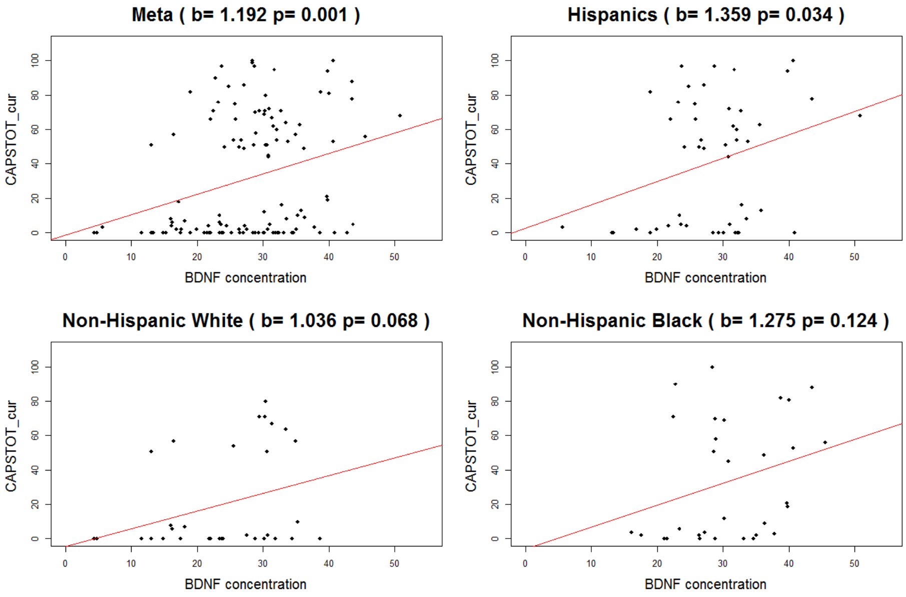


Figure 2.3 Scatter plot showing a positive correlation between BDNF blood concentration and PTSD severity.

Why only Hispanics seem to be susceptible by this polymorphism on BDNF needs to be further investigated. Particularly in light of the fact that several studies reported Hispanics have markedly higher risk level of developing PTSD and experience more severe symptoms compared to non-Hispanic Whites and non-Hispanic Blacks. These studies include military and civilian cohorts: in OEF/OIF veterans [32], in Vietnam War veterans [33], in police officers [34], after a terrorist attack [35], and after a natural disaster like a hurricane [36].

We plan to further investigate the following issues: (1) eQTL and mQTL analysis of identified SNPs, (2) genetic influence on metabolic profile and other intermediate molecular phenotypes (this kind of analysis helps fill in the mechanistic detail in between variants associated with disease phenotypes), (3) comparison of self-reported ethnicity/race and predictions from genetic admixture models, (4) replication study on publicly available dataset (dbGaP Study Accession: phs000864.v1.p1), and (5) fine mapping of local LD structure in the three ancestral subgroups.

Expression quantitative trait loci (eQTL) analysis to investigate the effects of DNA variants on gene expression and detecting their possible roles in developing PTSD

These top significant SNPs are not uniformly distributed across chromosomes. The histogram in Figure 2.4 demonstrates that the obtained SNPs are densely located in chromosomes 1 and 2.

We then tested if these SNPs are in Linkage Disequilibrium (LD). For each SNP, a 500 kb interval and a correlation threshold of 0.7 was considered. The SNPs in LD were then mapped to the genes hosting these SNPs (or the closest gene). These SNPs were almost all located inside a set of 141 unique genes which are mostly located in chromosomes 10 and 1 (**Figure 2.5**). Moreover, we tested to see if any of these SNPs on each unique chromosome are in LD with each other. Chromosome 4 contains the largest SNP pairs in LD where 80 SNPs have correlation coefficient over 0.7. Figure 2.6 represents the pairwise distances in Mb among the discovered SNPs across different chromosomes. Chromosome 4 contains 80 pairwise SNPs in LD, some of which are quite distant from each other, up to 0.4 Mb. Other chromosomes have fewer numbers of SNPs in LD, with LD pairs fairly close along the chromosome.

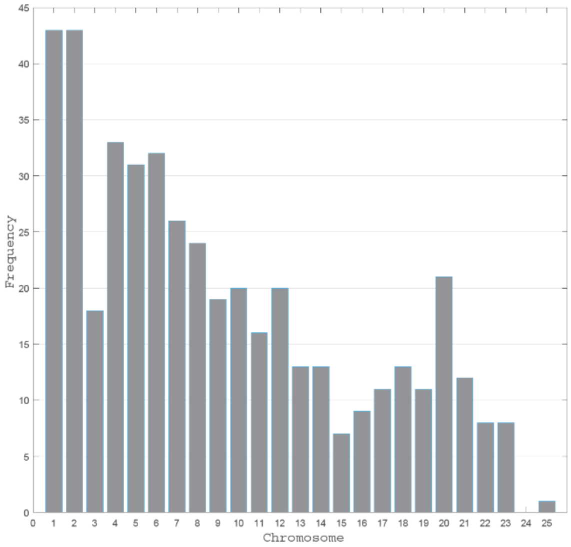


Figure 2.4 Distribution of the associated SNPs across chromosomes.

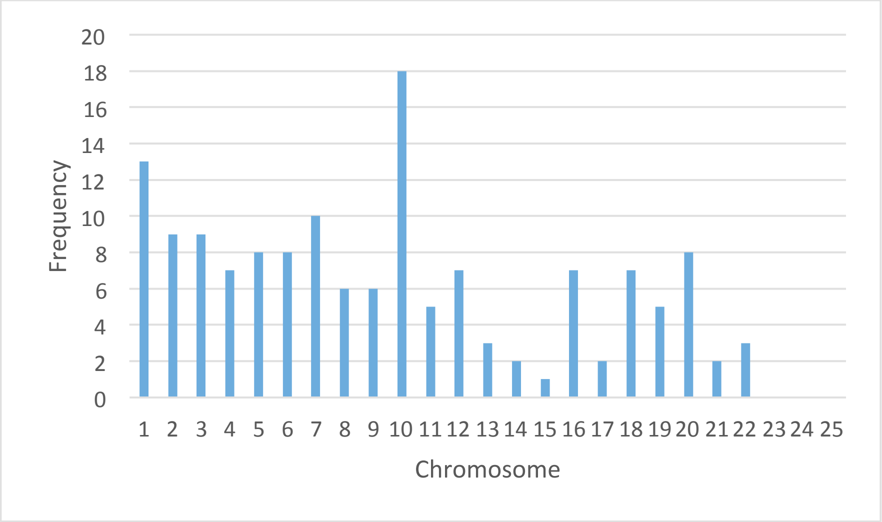


Figure 2.5 Distribution of the associated genes across chromosomes from GWAS.

According to Figure 2.7, except for chromosomes 6 and 9, the pairwise correlations between the detected SNPs on the other chromosomes are over 0.75. This is an indication of interconnection between the detected SNPs on each chromosome and reflects the fact that these SNPs mostly refer to a particular region on each chromosome. For instance, we showed that the number of the detected SNPs on chromosome 4 in LD is the largest among the entire detected SNPs and the median distance between the SNPs is 750 kb on chromosome 4 (the LD population is extracted from the hg19/1000 genome project results). The 33 SNPs located on chromosome 4 have 40 distinct pairs in LD with correlation over 0.6. The first three correlation quantiles are over 0.67. As an example, Figure 2.8 shows the locations of the detected SNPs on chromosome 4, based on their positions from 33 Mb to 50 Mb. Tightness of a group of SNPs is illustrated around 34 Mb, which is located in a non-coding region. SNP rs1435389 (p=2.77E-05) is located in a dense coding region but is not located inside a particular gene or in the 500 kb vicinity of other genes. However, rs10017276 (on the right hand side) is located inside CORIN.

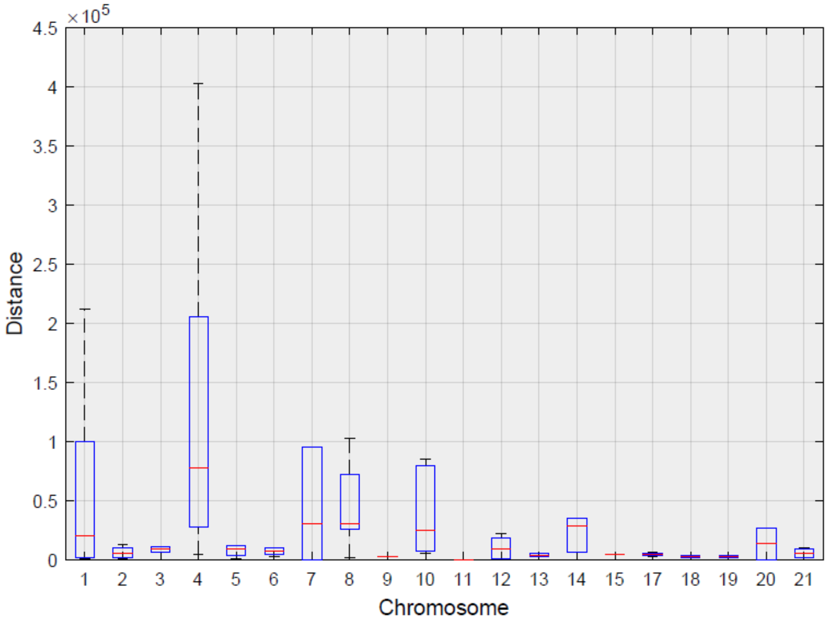


Figure 2.6 Distribution of pairwise distances between discovered SNPs across chromosomes.

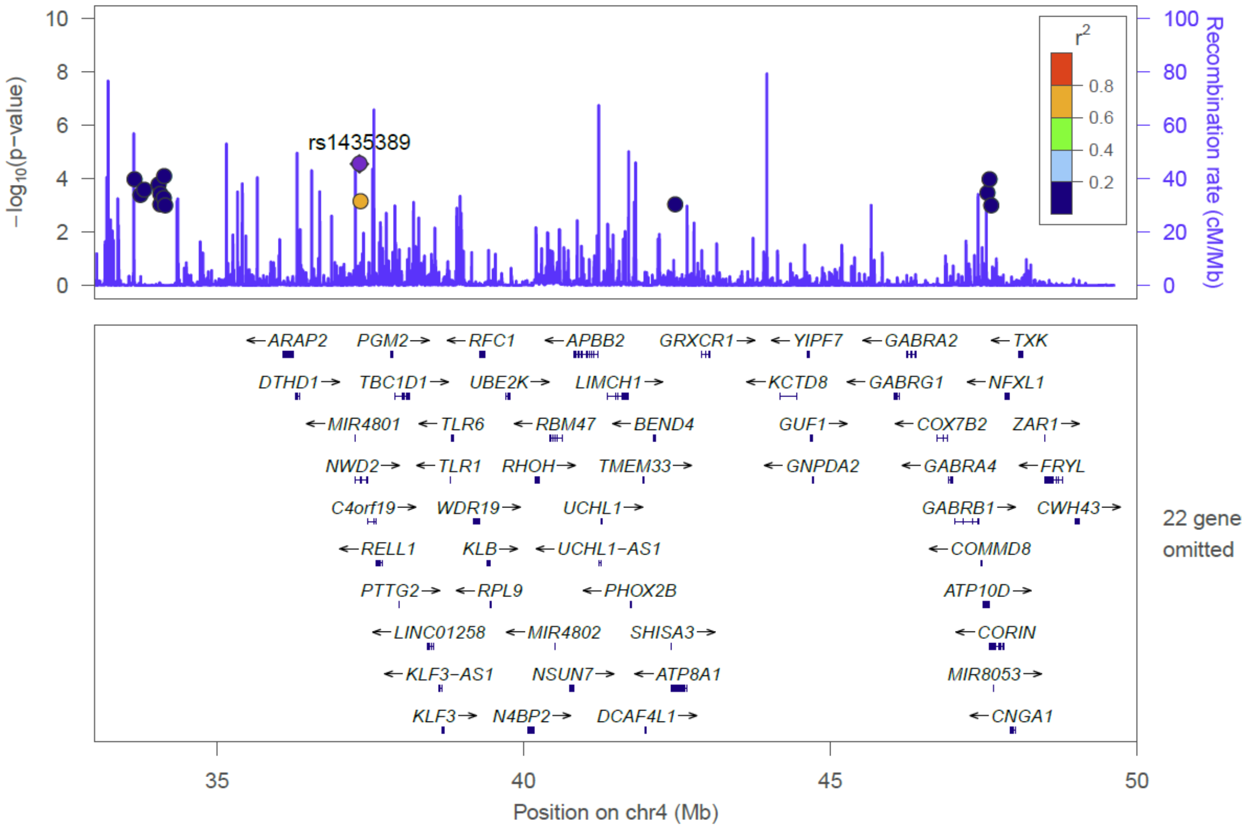


Figure 2.7 Pairwise correlation between the discovered SNPs across chromosomes.

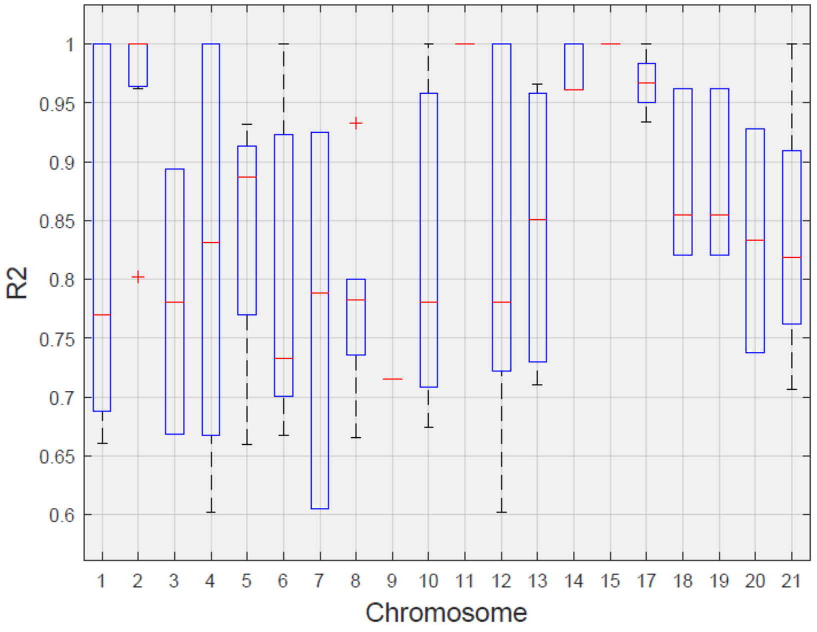


Figure 2.8 Location of detected SNPs on chromosome 4 from 33 to 50 Mb.

eQTL analysis was conducted using Matrix eQTL [37]. To do so, first the non-shared samples between the original SNP data and the mRNA data were removed. In total, 51 cases and 64 controls remained, which were all males. Then, genotypes were coded as a single allele dosage number to be used in Matrix eQTL. We have used Agilent platform to prepare the mRNA gene expression data. This dataset contains 46155 probes. 29005 probes were residing in non-coding regions and were removed from further analysis. 17150 remaining probes were located in coding regions. Linear additive models were used to test the interactions among the quantitative loci and expression of the genes. We did not limit our analysis to the gene level but performed interaction between the loci and the entire probes. mRNA data was batch corrected using the frozen Surrogate Variable Analysis (fSVA) method while controlling for Body Mass Index (BMI) and age. The two sided t-test was conducted to check which probes were differentially expressed between cases and controls. The distribution of the differentially expressed probes based on their p-values are depicted in Figure 2.9. In total, 82 probes were highly differentially expressed (p<0.01).

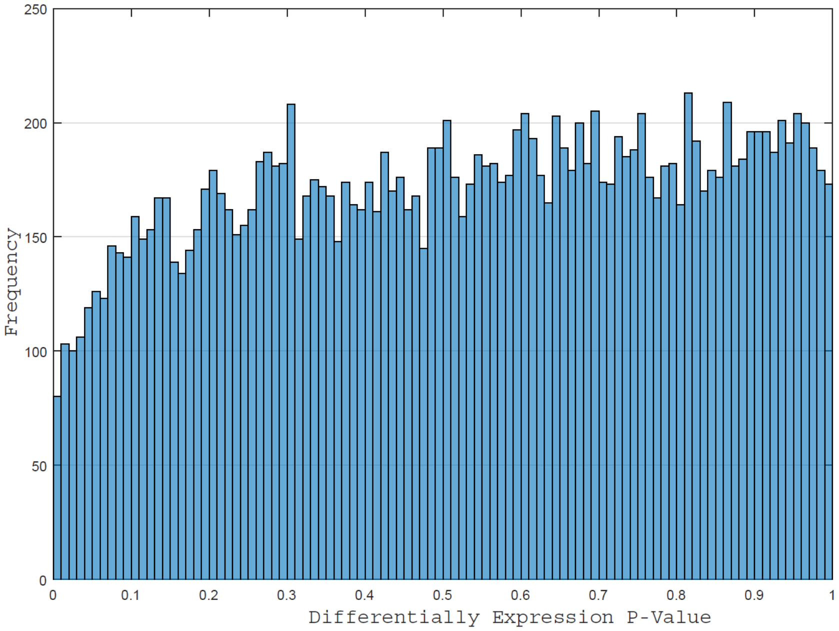


Figure 2.9 Histogram of the mRNA probe p-values between cases and controls.

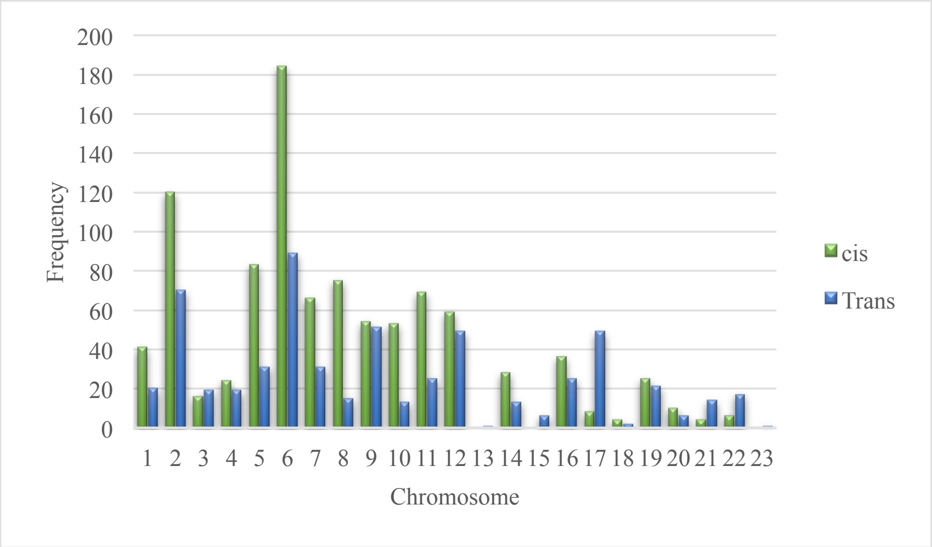


Figure 2.10 Histogram of cis and trans acting regulatory loci.

We have checked the cis and trans-acting eQTLs associated with expression of the 17150 gene expression probes. A 500 kb distance was taken as the look up window for cis-acting eQTLs. 965 cis-acting and 588 trans-acting eQTLs having p-values less than 0.001 were detected. Both cis and trans-regulatory eQTLs were mostly distributed on chromosome 6. The overall distribution of the detected eQTLs are represented in Figure 2.10.

The detected eQTLs correspond to 229 unique probes, corresponding to 221 unique genes. In order to further analyze the detected cis-acting eQTLs, we concentrated on the top eQTLs (p<1E-9). These eQTLs correspond to 48 expression probes mapped to 47 unique genes. We then performed disease association analysis. Three main neurologically-related disorders enriched with these genes are neural tube defects, shock disease, and nervous system impairment. For each single detected probe, we have extracted the corresponding SNPs along with the chromosomal location and the association p-value between the SNP and the probe. These three neurological-disorders associated chromosomal locations are presented in Tables 2.2-2.4.

Table 2.2 eQTLs highly associated with the transcripts and their corresponding chromosomal location related to Shock.

Gene	SNP	Probe	Association p-value	Location
FKBP1A	rs6041750	A_23_P397238	0.000536	20p13
CRYBB2	rs9612371	A_23_P425066	0.000189	22q11.23
	rs107017	A_23_P425066	0.000646	
	rs16997431	A_23_P425066	0.000674	
	rs6003692	A_23_P425066	0.000519	
PRDX2	rs1205170	A_24_P168416	0.000792	19p13.13
AHSA2	rs4671401	A_23_P372467	5.48E-06	2p15
	rs2290324	A_23_P372467	1.53E-05	
	rs1809028	A_23_P372467	2.81E-05	
SRI	rs1063964	A_23_P59718	1.25E-12	7q21.12

Table 2.3 eQTLs highly associated with the transcripts and their corresponding chromosomal location related nervous system impairments.

Gene	SNP	Probe	Association p-value	Location
GALC	rs378816	A_23_P25964	6.91E-07	7q36.1
	rs405567	A_23_P25964	1.87E-06	
	rs3213917	A_23_P25964	1.94E-05	
	rs398607	A_23_P25964	6.91E-07	
PEX6	rs9381225	A_23_P42144	4.35E-06	6p21.1
	rs2234185	A_23_P42144	4.44E-06	
	rs2007950	A_23_P42144	5.76E-06	
	rs6941212	A_23_P42144	0.000162	
	rs13199873	A_23_P42144	0.000251	
	rs1129187	A_23_P42144	0.000279	
	rs3763236	A_23_P42144	0.000675	
	rs2395943	A_23_P42144	1.66E-09	
PPT1	rs10889147	A_23_P62659	0.00053	1p34.2
VAPB	rs2268920	A_23_P91293	0.000412	20q13.32
	rs9679935	A_23_P91293	3.07E-05	
IGHMBP2	rs546382	A_23_P393713	2.92E-06	11q13.2
	rs660614	A_23_P393713	1.42E-05	
	rs629426	A_23_P393713	1.42E-05	
	rs636049	A_23_P393713	3.02E-05	
	rs619727	A_23_P393713	5.88E-05	
	rs604524	A_23_P393713	0.000627	
	rs622082	A_23_P393713	2.92E-06	
SUMF1	rs4685744	A_23_P69242	0.000961	3p26.2
LMNA	rs6682411	A_23_P34835	0.000357	1q22
	rs6427085	A_23_P34835	0.000589	
	rs11264336	A_23_P34835	0.000168	
MFN2	rs2295281	A_23_P126135	3.73E-05	1p36.22
MTRR	rs162036	A_23_P252211	3.36E-09	5p15.31
	rs12347	A_23_P252211	3.36E-09	
	rs9332	A_23_P252211	3.36E-09	
	rs2640658	A_23_P252211	3.89E-07	
	rs10380	A_23_P252211	1.30E-05	
	rs161871	A_23_P252211	1.31E-05	
	rs16879410	A_23_P252211	3.19E-05	
	rs3733784	A_23_P252211	6.34E-05	
	rs16879305	A_23_P252211	6.74E-05	
	rs1046014	A_23_P252211	0.000730	

Table 2.4 cis-eQTLs highly associated with the transcripts and their corresponding chromosomal location related to neural tube defects.

Gene	SNP	Probe	Association p-value	Location
LRRC6	rs1048490	A_23_P112004	1.08E-12	8q24.22
	rs3909640	A_23_P112004	1.25E-12	
	rs10216529	A_23_P112004	6.72E-12	
	rs7841637	A_23_P112004	1.08E-07	
	rs853308	A_23_P112004	4.43E-06	
	rs2739024	A_23_P112004	5.58E-06	
	rs2280871	A_23_P112004	7.67E-06	
	rs3843562	A_23_P112004	1.11E-05	
	rs4480107	A_23_P112004	2.46E-05	
	rs853321	A_23_P112004	4.91E-05	
	rs2052701	A_23_P112004	5.15E-05	
	rs1469263	A_23_P112004	0.000116	
	rs11996730	A_23_P112004	0.000139	
	rs4736611	A_23_P112004	0.000162	
	rs2272681	A_23_P112004	0.000198	
	rs2293979	A_23_P112004	0.000198	
	rs7004199	A_23_P112004	0.000627	
	rs11988034	A_23_P112004	1.08E-12	
MTRR	rs12347	A_23_P252211	3.36E-09	5p15.31
	rs9332	A_23_P252211	3.36E-09	
	rs2640658	A_23_P252211	3.89E-07	
	rs10380	A_23_P252211	1.30E-05	
	rs161871	A_23_P252211	1.31E-05	
	rs16879410	A_23_P252211	3.19E-05	
	rs3733784	A_23_P252211	6.34E-05	
	rs16879305	A_23_P252211	6.74E-05	
	rs1046014	A_23_P252211	0.000730	

2.2 Analysis of recalled PTSD subjects for markers and predictors of symptom changes

35 and 33 of the 83 PTSD and 83 Control subjects from the Original Biomarkers Study returned for follow-up evaluation approximately 2-3 years following their original enrollment in the study. Many of the PTSD subjects experienced a significant change in symptoms over this period. We have used the available molecular and clinical data from these two time points to identify molecular signals associated with symptom change, or signals which can predict future symptom change. A summary of the available recall data is shown in Table 2.5.

Table 2.5 Summary of available clinical and molecular data for recalled subjects at Time 1 (T1) and Time 2 (T2).

	# of PTSD Subjects from T1	# of Control Subjects from T1	# of PTSD Subjects from T2	# of Control Subjects from T2
Clinical	35	33	35	33
CLIA Lab	34	33	29	25
Endocrine	34	33	29	25
DNA Methylation	33	33	14	8
Metabolite	35	33	0	0
Protein	34	32	0	0

For data types with molecular data at both T1 and T2 (endocrine, CLIA Lab), we searched for features that were associated with changes in symptoms. Specifically, we identified variables whose change between T1 and T2 was correlated with the change in CAPS Total Current scores between the same time points. One endocrine marker, 5-alpha-reductase, and three CLIA Lab markers were significantly correlated with ΔCAPS ($p<0.01$), indicating that they could be used to track symptom recovery in PTSD subjects. Additionally, none of the CLIA Lab markers correlated with symptom change were previously identified as biomarkers of PTSD ($p>0.01$ in 83-83 Discovery cohort). Based on this cohort, these markers cannot be used to predict the magnitude of PTSD symptoms (or distinguish PTSD from Controls), but can be used to determine symptom changes over time. 5-alpha-reductase was also correlated with symptoms changes, but was also significantly different between PTSD and Control subjects at T1. However, though a significant difference in means exists, 5-alpha-reductase does not perform well as a diagnostic biomarker in the 83-83 training dataset. Figure 2.11 illustrates the findings associated with 5-alpha-reductase.

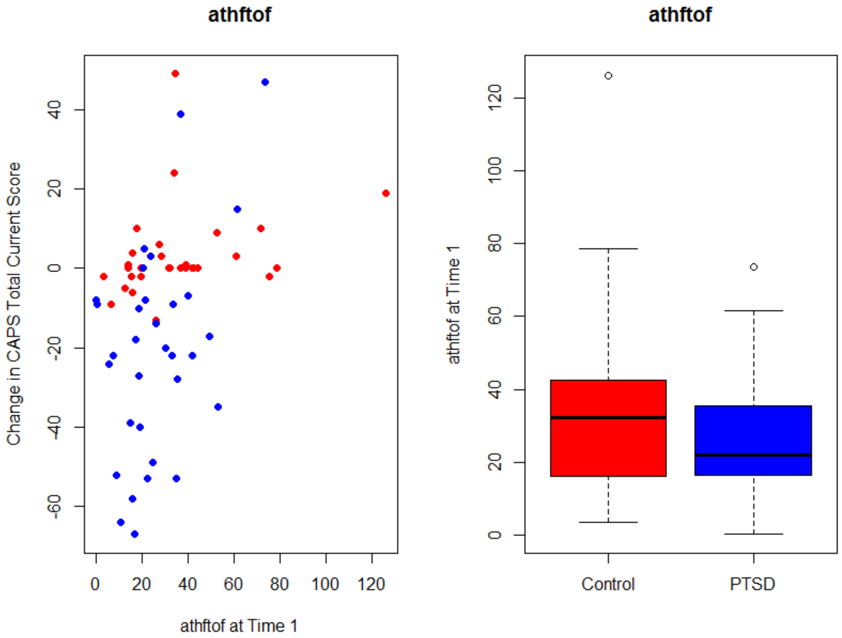


Figure 2.11 Overview of 5-alpha-reductase (athftof) signal in recalled subjects. Left: 5-alpha-reductase levels at T1 are correlated with ΔCAPS between T1 and T2 ($p<0.01$,FDR<0.1). Right: 5-alpha-reductase is also differentially expressed between Control (red) and PTSD (blue) subjects at T1 ($p<0.01$), though it does not perform well as a diagnostic marker (max AUC fitted on training data: 0.528).

In molecular data types without the completed recall subjects at T2, we identified molecular markers at T1 that can predict future symptom change. Table 2.6 summarizes the results of this analysis for all available data types. An example result from the metabolite dataset is shown in Figure 2.12, illustrating the correlation between T1 C-glycosyltryptophan levels and future CAPS changes. Similar to the CLIA lab results, C-glycosyltryptophan is not differentially expressed (p>0.01) between PTSD and Controls at T1, indicating is not able to distinguish the magnitude of PTSD symptoms, but instead is predictive of future symptom changes.

Table 2.6 Summary of significant predictive features.

Data Type	# of features considered	# of features predictive of future ΔCAPS (p<0.01)	# of features predictive of future ΔCAPS (FDR<0.01)
DNA Methylation	429948	11514	6
mRNA	50599	324	0
Metabolite	416	4	1
Protein	96	0	0
miRNA	43	0	0
Endocrine	35	1	1
CLIA Lab	44	0	0

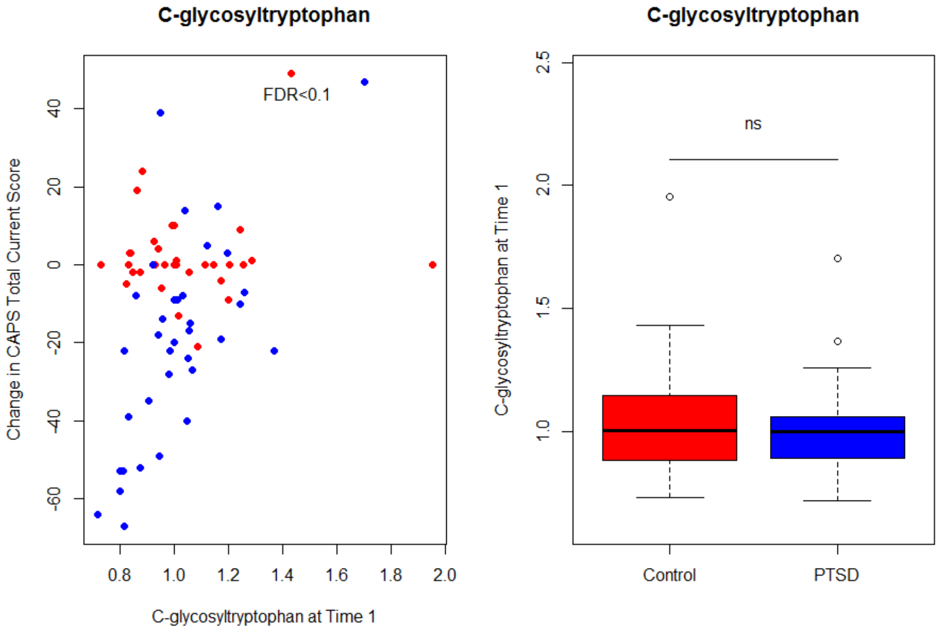


Figure 2.12 Overview of C-glycosyltryptophan signals in recalled subjects. Left: Normalized levels of C-glycosyltryptophan are correlated with ΔCAPS between T1 and T2 (p<0.01 and FDR<0.1). Right: C-glycosyltryptophan levels are not significantly different between Controls (red) and PTSD (blue) at T1 (p>0.01).

2.3 Application of Weighted Gene Correlation Network Analysis to PTSD DNA Methylation dataset

Weighted Gene Correlation Network Analysis (WGCNA) is an unsupervised framework for discovering networks of genes, modules, which are co-expressed but not necessary differentially expressed within a cohort [38]. We applied WGCNA to the 81/82 group (group 1) and validated separately on the 40/29 group (group 2) of PTSD subjects, performing functional enrichment analysis and module-trait association on each. The output of WGCNA on DNA methylation data is a set of modules where all of the CpG sites within the module are co-expressed. Due to the large size of the probes, we focus only on the differentially methylated set of CpG sites across the modules. From this, we observe several potential biomarkers and important pathways.

We show the discovered CpG sites and their respective module grouping in Table 2.7 for group 1. Functional enrichment analysis of the discovered modules using DAVID [39-40] and WebGestalt [41] are shown in Figure 2.13 for both groups. Of interest here are the olfactory transduction and taste transduction pathways as well as several neurological impairments, which are reproduced in both the 81/82 group and the 40/29 group. The independent identification of these pathways in both groups hint at the possible importance of these to PTSD. In addition, this is in accordance with a study by Chen et al. [42] that finds 8 olfactory related genes to be upregulated with PTSD.

Table 2.7 Differentially Methylated CpG sites in each module. We focus our analysis on looking at the differentially methylated CpG within each module as well as the gene with which they are associated.

Module	p-value	CpG	Gene	Module	p-value	CpG	Gene
1	0.000749	cg26613312	RNF152	1	0.00816	cg26306869	S100A3
1	0.001142	cg27065979	NEK3	1	0.008743	cg24925163	SFT2D3
1	0.001614	cg24441911	RBP5	1	0.00963	cg27327475	WHSC1
1	0.002	cg24424217	ZNF511	1	0.009791	cg24866407	SLC17A9
1	0.00259	cg19287591	ULK3	1	0.009906	cg19869608	ANKRD52
1	0.002606	cg27638115	MALL	1	0.009954	cg18479961	MIR671
1	0.002643	cg24905370	NSA2	2	0.000976	cg27422507	DSG2
1	0.002662	cg23501292	LIG1	2	0.006762	cg21898527	TMEM30C
1	0.002994	cg16320885	LCE5A	2	0.007414	cg18646864	ORC1L
1	0.003446	cg25141818	SLC26A9	2	0.007594	cg25422051	C11orf68
1	0.003687	cg26010099	S100A14	2	0.008507	cg27030081	ODF1
1	0.004063	cg18146398	CCR1	3	0.00119	cg27259271	SDCCAG8
1	0.004336	cg26722179	TMEM175	3	0.001612	cg21169285	MAPK9
1	0.004681	cg26514117	MIR181D	3	0.002442	cg27586378	SULT2B1
1	0.00505	cg18393023	SERPIND1	3	0.002554	cg26978776	CUZD1
1	0.00511	cg22289360	GCNT1	3	0.003714	cg27533700	EXOG
1	0.005412	cg22981296	RASD1	3	0.00536	cg26178664	CALML3
1	0.005487	cg20630690	ZGLP1	3	0.007625	cg23959009	ENOPH1
1	0.005586	cg27570233	FOXK2	4	0.000898	cg24254937	E2F2
1	0.005745	cg20388168	CUL9	4	0.007218	cg26591162	SRL
1	0.005985	cg25781926	MAPRE3	4	0.008304	cg14219236	SNORD114-22
1	0.006222	cg22923050	DDX49	4	0.009671	cg20948262	FRRS1
1	0.006402	cg25635303	DHX33	5	0.001199	cg26736540	TFAP2C
1	0.006526	cg26539468	CCDC111	6	0.000231	cg21316772	B4GALT1
1	0.007531	cg03569637	LOC100233209	6	0.006741	cg26227957	KIAA0090
1	0.007684	cg26218982	LACRT	7	0.000989	cg26589669	GSTO2
1	0.007858	cg23686278	XRRA1	7	0.005963	cg10432569	MIR196A1

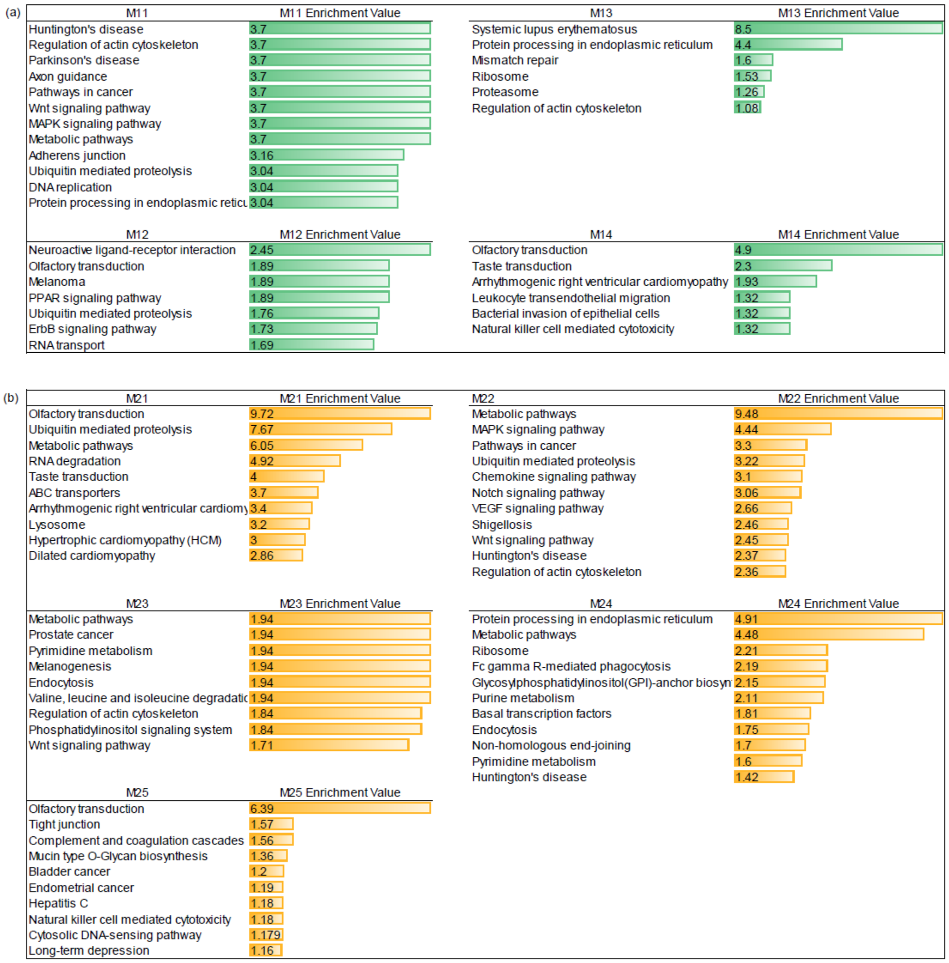
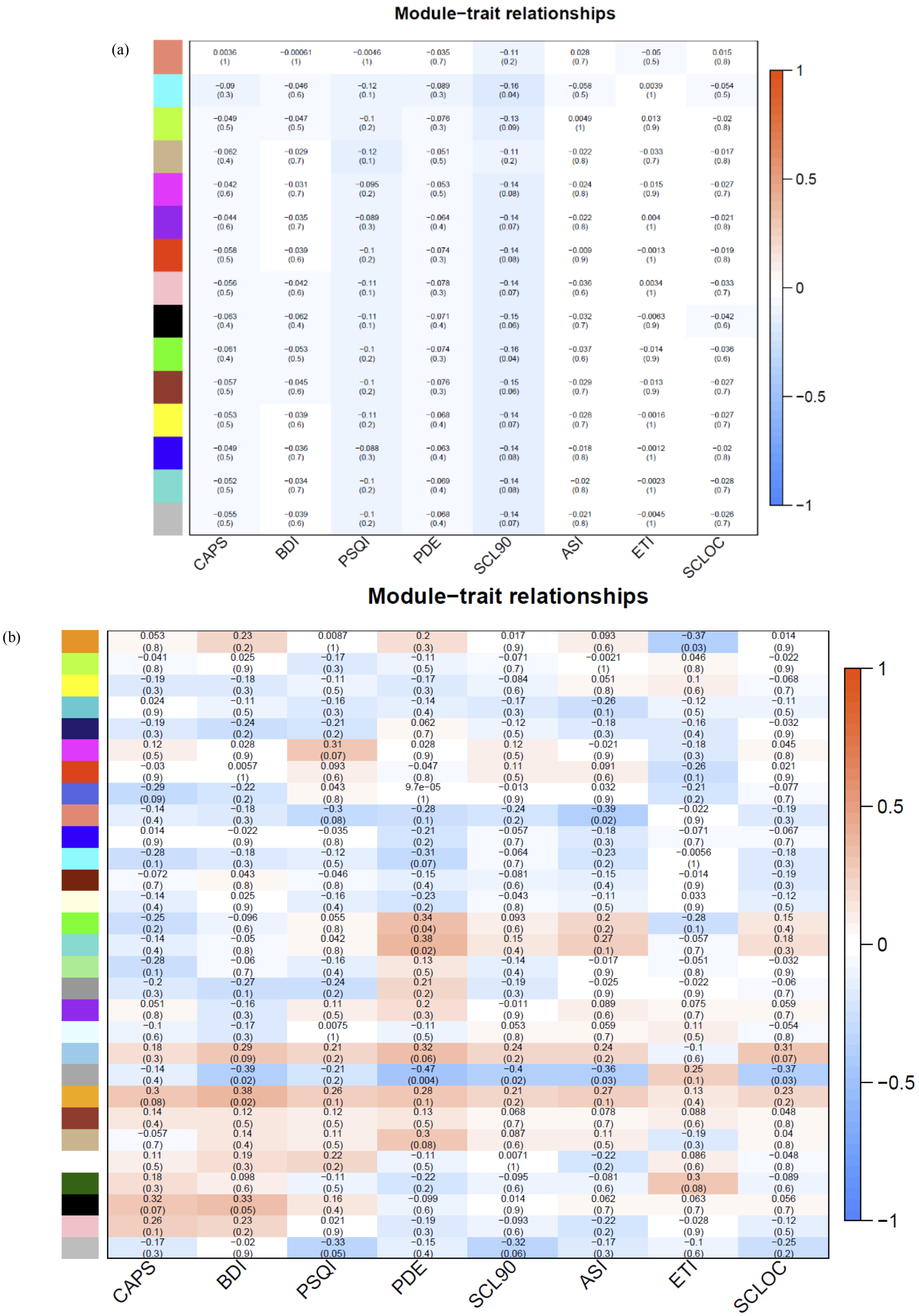


Figure 2.13 Activated Pathways of significant module in both groups. (a) Group 1 (b) Group 2. Functional enrichment analysis on the modules for both groups shows some consistency, particularly in olfactory transduction and taste transduction as well as a few neurological impairments.

To analyze the relevance of these modules to clinical traits and to further pinpoint specific important gene candidates, we calculate and show the correlation of each module to the clinical traits in Figure 2.14. Few of these modules are statistically significant to the clinical traits. However, if we look at the correlation between the differentially methylated genes from Table 2.7, we see stronger values. These results are shown in shown in Figure 2.15 and serve to validate the possible importance of some of these differentially methylated, module genes. Of note are the genes LCE5A, C11orf68, CALML3, SRL, and MIR196A1, but also highly correlated with clinical symptoms.



(a)

CAPS	BDI	PSQI	PDE	SCL90	ASI	ETI	SCLOC
0.0036 (1)	−0.00061 (1)	−0.0046 (1)	−0.035 (0.7)	−0.11 (0.2)	0.028 (0.7)	−0.05 (0.5)	0.015 (0.8)
−0.09 (0.3)	−0.046 (0.6)	−0.12 (0.1)	−0.089 (0.3)	−0.16 (0.04)	−0.058 (0.5)	0.0039 (1)	−0.054 (0.5)
−0.049 (0.5)	−0.047 (0.5)	−0.1 (0.2)	−0.076 (0.3)	−0.13 (0.09)	0.0049 (1)	0.013 (0.9)	−0.02 (0.8)
−0.062 (0.4)	−0.029 (0.7)	−0.12 (0.1)	−0.051 (0.5)	−0.11 (0.2)	−0.022 (0.8)	−0.033 (0.7)	−0.017 (0.8)
−0.042 (0.6)	−0.031 (0.7)	−0.095 (0.2)	−0.053 (0.5)	−0.14 (0.08)	−0.024 (0.8)	−0.015 (0.9)	−0.027 (0.7)
−0.044 (0.6)	−0.035 (0.7)	−0.089 (0.3)	−0.064 (0.4)	−0.14 (0.07)	−0.022 (0.8)	0.004 (1)	−0.021 (0.8)
−0.058 (0.5)	−0.039 (0.6)	−0.1 (0.2)	−0.074 (0.3)	−0.14 (0.08)	−0.009 (0.9)	−0.0013 (1)	−0.019 (0.8)
−0.056 (0.5)	−0.042 (0.6)	−0.11 (0.1)	−0.078 (0.3)	−0.14 (0.07)	−0.036 (0.6)	0.0034 (1)	−0.033 (0.7)
−0.063 (0.4)	−0.062 (0.4)	−0.11 (0.1)	−0.071 (0.4)	−0.15 (0.06)	−0.032 (0.7)	−0.0063 (0.9)	−0.042 (0.6)
−0.061 (0.4)	−0.053 (0.5)	−0.1 (0.2)	−0.074 (0.3)	−0.16 (0.04)	−0.037 (0.6)	−0.014 (0.9)	−0.036 (0.6)
−0.057 (0.5)	−0.045 (0.6)	−0.1 (0.2)	−0.076 (0.3)	−0.15 (0.06)	−0.029 (0.7)	−0.013 (0.9)	−0.027 (0.7)
−0.053 (0.5)	−0.039 (0.6)	−0.11 (0.2)	−0.068 (0.4)	−0.14 (0.07)	−0.028 (0.7)	−0.0016 (1)	−0.027 (0.7)
−0.049 (0.5)	−0.036 (0.7)	−0.088 (0.3)	−0.063 (0.4)	−0.14 (0.08)	−0.018 (0.8)	−0.0012 (1)	−0.02 (0.8)
−0.052 (0.5)	−0.034 (0.7)	−0.1 (0.2)	−0.069 (0.4)	−0.14 (0.08)	−0.02 (0.8)	−0.0023 (1)	−0.028 (0.7)
−0.055 (0.5)	−0.039 (0.6)	−0.1 (0.2)	−0.068 (0.4)	−0.14 (0.07)	−0.021 (0.8)	−0.0045 (1)	−0.026 (0.7)

(b)

CAPS	BDI	PSQI	PDE	SCL90	ASI	ETI	SCLOC
0.053 (0.8)	0.23 (0.2)	0.0087 (1)	0.2 (0.3)	0.017 (0.9)	0.093 (0.6)	−0.37 (0.03)	0.014 (0.9)
−0.041 (0.8)	0.025 (0.9)	−0.17 (0.3)	−0.11 (0.5)	−0.071 (0.7)	−0.0021 (1)	0.046 (0.8)	−0.022 (0.9)
−0.19 (0.3)	−0.18 (0.3)	−0.11 (0.5)	−0.17 (0.3)	−0.084 (0.6)	0.051 (0.8)	0.1 (0.6)	−0.068 (0.7)
0.024 (0.9)	−0.11 (0.5)	−0.16 (0.3)	−0.14 (0.4)	−0.17 (0.3)	−0.26 (0.1)	−0.12 (0.5)	−0.11 (0.5)
−0.19 (0.3)	−0.24 (0.2)	−0.21 (0.2)	0.062 (0.7)	−0.12 (0.5)	−0.18 (0.3)	−0.16 (0.4)	−0.032 (0.9)
0.12 (0.5)	0.028 (0.9)	0.31 (0.07)	0.028 (0.9)	0.12 (0.5)	−0.021 (0.9)	−0.18 (0.3)	0.045 (0.8)
−0.03 (0.9)	0.0057 (1)	0.093 (0.6)	−0.047 (0.8)	0.11 (0.5)	0.091 (0.6)	−0.26 (0.1)	0.021 (0.9)
−0.29 (0.09)	−0.22 (0.2)	0.043 (0.8)	9.7e−05 (1)	−0.013 (0.9)	0.032 (0.9)	−0.21 (0.2)	−0.077 (0.7)
−0.14 (0.4)	−0.18 (0.3)	−0.3 (0.08)	−0.28 (0.1)	−0.24 (0.2)	−0.39 (0.02)	−0.022 (0.9)	−0.19 (0.3)
0.014 (0.9)	−0.022 (0.9)	−0.035 (0.8)	−0.21 (0.2)	−0.057 (0.7)	−0.18 (0.3)	−0.071 (0.7)	−0.067 (0.7)
−0.28 (0.1)	−0.18 (0.3)	−0.12 (0.5)	−0.31 (0.07)	−0.064 (0.7)	−0.23 (0.2)	−0.0056 (1)	−0.18 (0.3)
−0.072 (0.7)	0.043 (0.8)	−0.046 (0.8)	−0.15 (0.4)	−0.081 (0.6)	−0.15 (0.4)	−0.014 (0.9)	−0.19 (0.3)
−0.14 (0.4)	0.025 (0.9)	−0.16 (0.4)	−0.23 (0.2)	−0.043 (0.8)	−0.11 (0.5)	0.033 (0.9)	−0.12 (0.5)
−0.25 (0.2)	−0.096 (0.6)	0.055 (0.8)	0.34 (0.04)	0.093 (0.6)	0.2 (0.2)	−0.28 (0.1)	0.15 (0.4)
−0.14 (0.4)	−0.05 (0.8)	0.042 (0.8)	0.38 (0.02)	0.15 (0.4)	0.27 (0.1)	−0.057 (0.7)	0.18 (0.3)
−0.28 (0.1)	−0.06 (0.7)	−0.16 (0.4)	0.13 (0.5)	−0.14 (0.4)	−0.017 (0.9)	−0.051 (0.8)	−0.032 (0.9)
−0.2 (0.3)	−0.27 (0.1)	−0.24 (0.2)	0.21 (0.2)	−0.19 (0.3)	−0.025 (0.9)	−0.022 (0.9)	−0.06 (0.7)
0.051 (0.8)	−0.16 (0.3)	0.11 (0.5)	0.2 (0.3)	−0.011 (0.9)	0.089 (0.6)	0.075 (0.7)	0.059 (0.7)
−0.1 (0.6)	−0.17 (0.3)	0.0075 (1)	−0.11 (0.5)	0.053 (0.8)	0.059 (0.7)	0.11 (0.5)	−0.054 (0.8)
0.18 (0.3)	0.29 (0.09)	0.21 (0.2)	0.32 (0.06)	0.24 (0.2)	0.24 (0.2)	−0.1 (0.6)	0.31 (0.07)
−0.14 (0.4)	−0.39 (0.02)	−0.21 (0.2)	−0.47 (0.004)	−0.4 (0.02)	−0.36 (0.03)	0.25 (0.1)	−0.37 (0.03)
0.3 (0.08)	0.38 (0.02)	0.26 (0.1)	0.28 (0.1)	0.21 (0.2)	0.27 (0.1)	0.13 (0.4)	0.23 (0.2)
0.14 (0.4)	0.12 (0.5)	0.12 (0.5)	0.13 (0.5)	0.068 (0.7)	0.078 (0.7)	0.088 (0.6)	0.048 (0.8)
−0.057 (0.7)	0.14 (0.4)	0.11 (0.5)	0.3 (0.08)	0.087 (0.6)	0.11 (0.5)	−0.19 (0.3)	0.04 (0.8)
0.11 (0.5)	0.19 (0.3)	0.22 (0.2)	−0.11 (0.5)	0.0071 (1)	−0.22 (0.2)	0.086 (0.6)	−0.048 (0.8)
0.18 (0.3)	0.098 (0.6)	−0.11 (0.5)	−0.22 (0.2)	−0.095 (0.6)	−0.081 (0.6)	0.3 (0.08)	−0.089 (0.6)
0.32 (0.07)	0.33 (0.05)	0.16 (0.4)	−0.099 (0.6)	0.014 (0.9)	0.062 (0.7)	0.063 (0.7)	0.056 (0.7)
0.26 (0.1)	0.23 (0.2)	0.021 (0.9)	−0.19 (0.3)	−0.093 (0.6)	−0.22 (0.2)	−0.028 (0.9)	−0.12 (0.5)
−0.17 (0.3)	−0.02 (0.9)	−0.33 (0.05)	−0.15 (0.4)	−0.32 (0.06)	−0.17 (0.3)	−0.1 (0.6)	−0.25 (0.2)

Figure 2.14 Correlations between each module and the clinical traits. (a) Group 1 (b) Group 2. Correlation was found by calculating the first module eigenvector and correlating it with the clinical values. Only a few modules are significantly correlated.

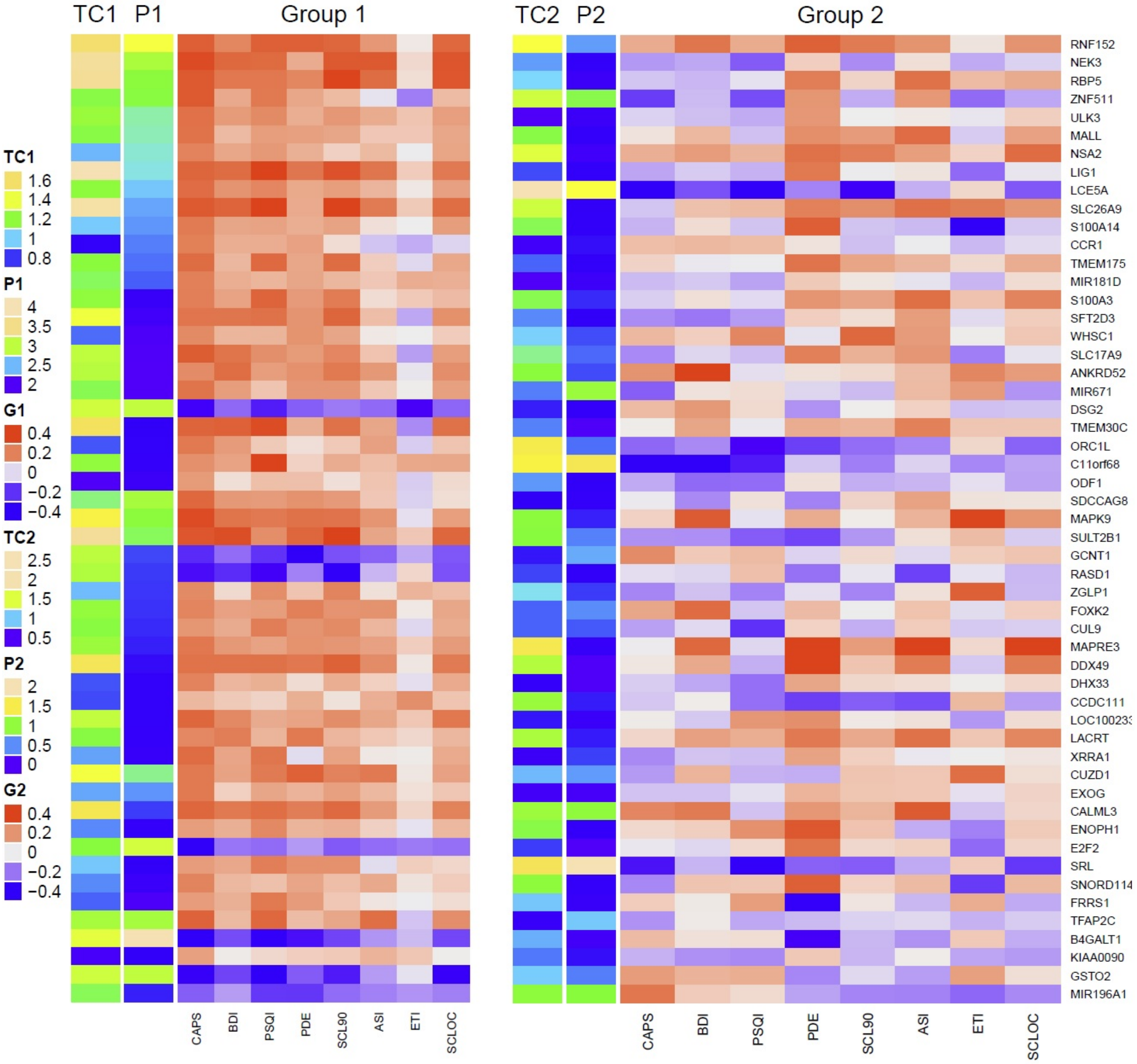


Figure 2.15 Correlations between the differentially methylated genes in Table 2.7 and the clinical traits. TC (Total Correlation) is the sum of the correlations for each gene and P is the -Log$_{10}$(p-value) of the gene between cases and controls. Most of these genes are strongly correlated with at least 2 clinical traits, indicating a possibility of importance in PTSD.

3. Application and development of classification and biomarker identification tools to finalize biomarker panels in single ‘omic datasets

3.1 Data-driven biomarker discovery and classification using soft computing techniques

In this work, we have extended the soft computing approach proposed in [43] by increasing the computational robustness and noise resistance abilities of the algorithm by employing Interval Type-2 (IT2) fuzzy sets instead of ordinary fuzzy membership functions. To do so, we have added a type-reduction module to the algorithm in order to adapt the type-2 fuzzy part with the other steps of the algorithm. The general structure of IT2 fuzzy systems is provided in Figure 3.1. The overall pipeline of this approach is represented in Figure 3.2.

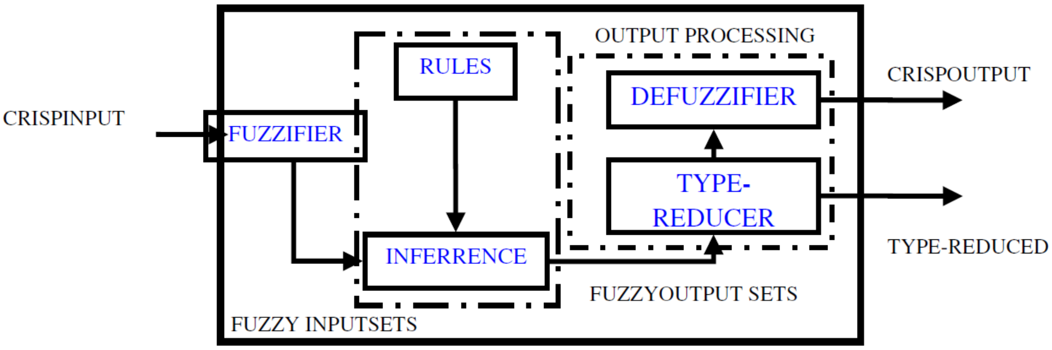


Figure 3.1 A schematic view of IT2 fuzzy systems.

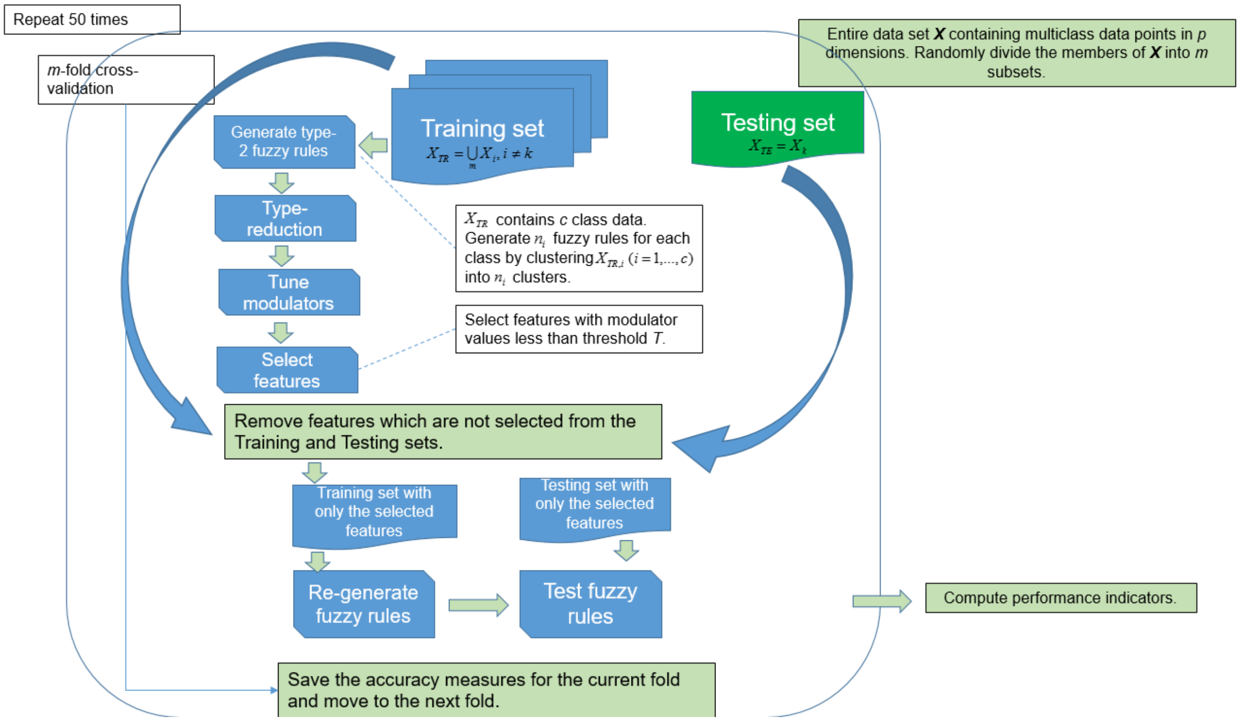

Repeat 50 times
m-fold cross-
validation
Entire data set ***X*** containing multiclass data points in *p* dimensions. Randomly divide the members of ***X*** into *m* subsets.
Training set
$X_{TR} = \bigcup_{m} X_{i}, i \neq k$
Testing set
$X_{TE} = X_{k}$
Generate type-
2 fuzzy rules
Type-
reduction
Tune
modulators
Select
features
X_{TR} contains *c* class data.
Generate n_i fuzzy rules for each
class by clustering $X_{TR,i}$ $(i = 1,...,c)$
into n_i clusters.
Select features with modulator
values less than threshold *T*.
Remove features which are not selected from the
Training and Testing sets.
Training set with
only the selected
features
Testing set with
only the selected
features
Re-generate
fuzzy rules
Test fuzzy
rules
Compute performance indicators.
Save the accuracy measures for the current fold
and move to the next fold.

Figure 3.2 Flowchart of the extended soft computing classification and biomarker discovery approach.

Application of soft computing techniques to PTSD mRNA dataset

We have implemented this algorithm on the Agilent 52-52 mRNA data. For comparison, we have also run Nearest Shrunken Centroids (NSC) classification and Logistic Regression with L1-regularization. The computational performance metrics along with the identified biomarkers are summarized in Tables 3.1-3.2.

Table 3.1. Classification performance measures for 52-52 mRNA data using extended fuzzy rule-based system, Nearest Shrunken Centroids and L1-Logistic Regression.

AUC: Area Under the ROC Curve; ER: Error Rate; MCC: Matthew Correlation Coefficient; MSPE: Mean Squared Predicted Error; Youden: Youden Index; PPV: Positive Predictive Value; NPV: Negative Predictive Value; TPR: True Positive Rate; FPR: False Positive Rate.

Metric	Nearest Shrunken Centroids	Logistic Regression with L1-regularization	Fuzzy rule-based system
AUC	0.599	0.543	0.598
ER	0.514	0.405	0.404
MCC	-0.019	0.195	0.197
MSPE	0.248	0.753	0.760
Youden	0.289	0.240	0.239
PPV	0.500	0.571	0.591
NPV	0.480	0.625	0.610
TPR	0.316	0.667	0.685
FPR	0.333	0.474	0.430

Table 3.2 Identified mRNA biomarkers using the fuzzy rule-based approach.

ID	Symbol	Function	p-value	FDR
A_24_P135444	AMFR	Colorectal cancer-related gene.	0.000016	0.99970
A_23_P208013	ZNF407	Causes syndromic intellectual disability.	0.000463	0.99970
A_33_P3356877	OR13C3	Olfactory receptors to initiate neural response that triggers the perception of smell.	0.001309	0.99970
A_33_P3401295	CRCT1	The encoded protein may be involved in amyotrophic lateral sclerosis and Parkinson's disease.	0.00189	0.99970
A_24_P90097	ADD3	Disease associated is dyscalculia A learning disability such difficulties as learning math concepts	0.001174	0.99970
A_33_P3280721	WLS	Skin related disorders.	0.001361	0.99970
A_33_P3221665	KIF21B	Inflammatory bowel disease.	0.001094	0.99970
A_23_P424900	C1orf88	to control cilia retraction as well as the liberation and duplication of the basal body/centrosome.	0.001029	0.99970
A_23_P105002	ROM1	Diseases associated with ROM1 include retinitis pigmentosa 7 and digenic and rom1-related retinitis pigmentosa.	0.000619	0.99970
A_24_P115507	SARDH	Diseases associated with SARDH include sarcosinemia and dimethylglycine dehydrogenase deficiency.	0.000558	0.99970
A_33_P3272260	SIRPB2	Lymphoma-related gene.	0.000183	0.99970
A_33_P3295415	ZBTB3	Colorectal cancer-related gene.	0.00163	0.99970
A_23_P61881	ARIH2	Can differentiate in the bone marrow into granulocytes.	0.001831	0.99970

Application of soft computing techniques to PTSD DNA Methylation dataset

Since the methylation dataset is very large, we decided to study subsets of probes in order to reduce the dimension of the search space. The choice of probes is motivated by biology of the disorder. The subset used were: 1) only the promoter region probes, 2) probes corresponding to genes that are expressed in brain and found in blood, and 3) promoter region probes of the genes from subset two. The second subset was curated by using the genes assayed in the Allen brain study for the adult and developing human brain [44-47]. Including the original dataset, we analyzed a total of four datasets (three subsets plus the full dataset). These are referred to as Promoter Only, Brain-All, Brian-Promoter and All, respectively. The available validation dataset consisted of 31 subjects in each PTSD and control groups. A subset of the validation data was separated to use for integration of various panels obtained by different methods. The validation performance was then computed on the remaining subset of validation data.
Using the extended fuzzy rule-based method mentioned above, we have implemented it on four splits of the Agilent 52-52 DNA methylation data: Promoter Only, Brain-All, Brain-Promoter, and All. The computational measures along with the identified biomarkers are represented in Tables 3.3-3.7.

Table 3.3 Extended fuzzy rule-based classification performance measures. Classification performance metrics on Agilent DNA methylation data. Predicted/Cross-validated performance measures (“P”) on the 52-52 cohort, and actual/validation performance measures (“A”) on 31-31 cohort.

Metric	**Blood (All)**		**Blood (Promoter)**		**Brain + Blood (All)**		**Brain + Blood (Promoter)**	
	P	A	P	A	P	A	P	A
AUC	0.6203± 0.1098	0.600	0.635 ± 0.1006	0.600	0.6301 ± 0.093	0.601	0.604 ± 0.1312	0.588
ER	0.40 ± 0.0872	0.439	0.396 ± 0.1008	0.453	0.389 ± 0.1001	0.408	0.398 ± 0.0951	0.422
PPV	0.603 ± 0.1237	0.532	0.599 ± 0.0946	0.551	0.612 ± 0.180	0.575	0.625 ± 0.1263	0.566
NPV	0.608 ± 0.1011	0.510	0.608 ± 0.1105	0.604	0.610 ± 0.1183	0.558	0.610 ± 0.1170	0.573
TPR	0.591 ± 0.0947	0.578	0.581 ± 0.1089	0.591	0.590 ± 0.1571	0.601	0.595 ± 0.1222	0.645
FPR	0.393 ± 0.0729	0.506	0.377 ± 0.089	0.519	0.374 ± 0.091	0.471	0.355 ± 0.0821	0.513

Table 3.4 Identified DNA Methylation biomarkers from All probes.

Probe ID	Gene Name	p-value
A_17_P16992190	KPNB1	2.64E-08
A_17_P15229454	WDR43	8.15E-08
A_17_P15441405	WNT5A	1.60E-07
A_17_P08336156	COL2A1	2.42E-07
A_17_P15491932	KY	3.79E-07
A_17_P15037255	CDC42	3.93E-07
A_17_P16442560	PACSIN3	4.20E-07
A_17_P16836061	HBA2	6.16E-07
A_17_P30223816	SKOR1	6.26E-07
A_17_P15113548	PDE4DIP	8.23E-07
A_17_P32480043	ZC4H2	9.35E-07
A_17_P16380506	SH3PXD2A	1.03E-06
A_17_P15051080	MACF1	1.15E-06
A_17_P16918077	GAN	1.64E-06
A_17_P32147249	NEFH	1.64E-06
A_17_P32154566	SELM	1.73E-06
A_17_P16562934	LRP1	1.79E-06
A_17_P16378261	LZTS2	1.85E-06
A_17_P15100937	LRIG2	1.89E-06
A_17_P02004261	KIF1A	1.95E-06
A_17_P22026036	FAM134A	2.49E-06
A_17_P09328713	DHRS7	2.49E-06
A_17_P16148119	OPLAH	2.50E-06
A_17_P07596187	EFCAB4A	3.42E-06
A_17_P16385788	CASP7	3.44E-06
A_17_P15281121	MAL	3.90E-06
A_17_P15554253	STIM2	4.23E-06
A_17_P16880097	TGFB1I1	4.30E-06
A_17_P15048443	KIAA0319L	4.97E-06
A_17_P27346623	RNU6ATAC	6.37E-06
A_17_P09929677	RUNDC2A	1.29E-05
A_17_P16877882	MAPK3	1.39E-05
A_17_P31564227	NUDT19	1.41E-05
A_17_P16462970	OVOL1	1.44E-05
A_17_P16836091	HBA1	1.45E-05
A_17_P16424413	MYOD1	1.60E-05
A_17_P06198312	SLCO5A1	1.64E-05
A_17_P27019833	ANKRD20A2	2.63E-05
A_17_P10947425	FFAR1	3.09E-05
A_17_P09929677	RUNDC2A	1.29E-05
A_17_P31564227	NUDT19	1.41E-05
A_17_P20899867	CNIH3	1.58E-05
A_17_P17286585	SREBF2	1.80E-05
A_17_P30831096	WSB1	2.42E-05
A_17_P31080544	BAHCC1	8.56E-06
A_17_P09150998	RASA3	8.82E-06
A_17_P25611800	MIR550A1	1.01E-05

Table 3.5 Identified DNA Methylation biomarkers from Brain-All probes.

Probe ID	Gene Name	p-value
A_17_P15441405	WNT5A	1.58E-07
A_17_P08336156	COL2A1	2.55E-07
A_17_P26986345	FANCG	2.77E-07
A_17_P15037255	CDC42	3.98E-07
A_17_P15055635	PLK3	4.53E-07
A_17_P26287785	PPP3CC	6.78E-07
A_17_P05262813	MICALL2	9.87E-07
A_17_P32889238	DGAT1	1.06E-06
A_17_P15051080	MACF1	1.07E-06
A_17_P16380506	SH3PXD2A	1.26E-06
A_17_P16079244	ENTPD4	1.68E-06
A_17_P16918077	GAN	1.76E-06
A_17_P02004261	KIF1A	1.87E-06
A_17_P16562934	LRP1	1.9E-06
A_17_P21225125	SIX2	2.07E-06
A_17_P09736379	GLCE	2.16E-06
A_17_P22026036	FAM134A	2.67E-06
A_17_P29597995	PCID2	3.13E-06
A_17_P16385788	CASP7	3.62E-06
A_17_P15403467	STK25	4.52E-06
A_17_P24938290	FBXO9	5.14E-06
A_17_P06807708	GADD45G	5.77E-06
A_17_P27457820	STAM	6.65E-06
A_17_P15136165	KCNJ9	7.57E-06
A_17_P21799473	DLX1	7.96E-06
A_17_P09150998	RASA3	8.75E-06
A_17_P00476260	NHLH2	9.43E-06
A_17_P23466228	PPP3CA	9.49E-06
A_17_P15500029	TSC22D2	9.75E-06
A_17_P17172157	PLCB4	9.99E-06
A_17_P16098330	NSMAF	1.02E-05
A_17_P15055788	HECTD3	1.18E-05
A_17_P20523826	TBX15	1.29E-05
A_17_P17209335	COL9A3	1.3E-05
A_17_P15778411	SERPINB6	1.35E-05
A_17_P00060115	EFHD2	1.41E-05
A_17_P16462970	OVOL1	1.42E-05
A_17_P16877882	MAPK3	1.5E-05
A_17_P16424413	MYOD1	1.58E-05
A_17_P16387276	GFRA1	1.8E-05
A_17_P15074726	USP33	1.91E-05
A_17_P17234558	SIM2	2.41E-05
A_17_P17213204	PCMTD2	2.54E-05
A_17_P16555982	DDN	2.72E-05
A_17_P16372461	TNKS2	2.87E-05
A_17_P16785654	EIF3J	2.88E-05
A_17_P16937597	UBE2G1	3.32E-05
A_17_P20784489	DDX59	4.4E-05
A_17_P17194852	DBNDD2	6.17E-05
A_17_P16292171	DHTKD1	6.39E-05
A_17_P16672133	SOX1	7.4E-05
A_17_P17236438	PSMG1	7.6E-05
A_17_P29827163	ACTN1	9.65E-05
A_17_P01232345	UGP2	4.15E-05
A_17_P17097970	TLE2	3.04E-05

Table 3.6 Identified DNA Methylation biomarkers from Brain-Promoter probes.

Probe ID	Gene Name	p-value
A_17_P26986345	FANCG	7.11E-07
A_17_P09155719	ANG	1.37E-06
A_17_P17028211	FASN	3.4E-06
A_17_P09736379	GLCE	4.59E-06
A_17_P21225125	SIX2	4.67E-06
A_17_P10625332	MAPRE2	1.35E-05
A_17_P23455530	EIF4E	3.23E-05
A_17_P16387276	GFRA1	3.55E-05
A_17_P31614741	FKRP	0.000047
A_17_P24102370	FST	5.26E-05
A_17_P27307280	PBX3	7.17E-05
A_17_P16966417	TIAF1	7.35E-05
A_17_P00777833	PPP2R5A	7.83E-05
A_17_P20784489	DDX59	7.89E-05
A_17_P16292171	DHTKD1	0.000106
A_17_P17194852	DBNDD2	0.000112
A_17_P15531898	DGKQ	0.000125
A_17_P15693436	PLK2	0.000127
A_17_P23676579	NR3C2	0.000205
A_17_P16079241	ENTPD4	0.000209
A_17_P10488594	NPTX1	0.000252
A_17_P15740601	ACSL6	0.000282
A_17_P16673810	MCF2L	0.000293
A_17_P16137330	ST3GAL1	0.000358
A_17_P16798818	DPP8	0.000433
A_17_P16683127	TEP1	0.000499
A_17_P15496101	CHST2	0.0005
A_17_P16096976	PENK	0.000516
A_17_P15834791	FAM46A	0.000523
A_17_P17279275	MCM5	0.000528
A_17_P10174699	FBXO31	0.000549
A_17_P22772943	TFDP2	0.000561
A_17_P31161966	SNRPD1	0.000259
A_17_P29056397	DTX1	0.000306
A_17_P17254685	DGCR14	0.000329
A_17_P30840930	TAOK1	0.000233
A_17_P04506430	FOXQ1	0.000201
A_17_P15831914	COL12A1	0.000192
A_17_P16992564	CDK5RAP3	0.000158
A_17_P15023754	FBXO44	0.000152
A_17_P05110439	TNFAIP3	0.000297
A_17_P16175582	PAX5	0.00037
A_17_P23839254	CDKN2AIP	0.000612
A_17_P16154930	JAK2	0.000616
A_17_P01670284	GAD1	0.000717
A_17_P28304520	LRRC32	0.00079
A_17_P20007295	MMP23B	0.000855
A_17_P15043445	EPB41	0.001014

Table 3.7 Identified DNA Methylation biomarkers from Promoter Only probes.

Probe ID	Gene Name	p-value
A_17_P15155447	RGS2	5.74E-08
A_17_P15441405	WNT5A	1.60E-07
A_17_P32480043	ZC4H2	9.35E-07
A_17_P17133179	GRAMD1A	9.85E-07
A_17_P21278637	CCDC85A	1.22E-06
A_17_P17028211	FASN	1.26E-06
A_17_P02004261	KIF1A	1.95E-06
A_17_P09736379	GLCE	2.17E-06
A_17_P16385788	CASP7	3.44E-06
A_17_P10625332	MAPRE2	6.39E-06
A_17_P15142174	ATP1B1	1.06E-05
A_17_P15752108	CDX1	1.48E-05
A_17_P16241713	PHF2	1.72E-05
A_17_P16228757	VPS13A	1.87E-05
A_17_P15971096	GTF2IRD1	2.31E-05
A_17_P30831096	WSB1	2.42E-05
A_17_P26032234	CCDC136	2.84E-05
A_17_P10947425	FFAR1	3.09E-05
A_17_P31437695	GALR1	3.35E-05
A_17_P20784489	DDX59	4.56E-05
A_17_P02902273	MSX1	4.57E-05
A_17_P16268167	PRDM12	5.31E-05
A_17_P10361837	NME1-NME2	5.69E-05
A_17_P32701000	MIR424	5.81E-05
A_17_P02942715	CPEB2	5.97E-05
A_17_P05778788	MIR183	2.58E-05
A_17_P21686874	EPC2	3.07E-06
A_17_P31614741	FKRP	2.39E-05
A_17_P15971096	GTF2IRD1	2.31E-05
A_17_P24938290	FBXO9	4.88E-06
A_17_P06807708	GADD45G	5.47E-06
A_17_P01615701	LY75-CD302	5.58E-06
A_17_P17026267	ACTG1	6.08E-06
A_17_P01623159	TBR1	6.16E-06
A_17_P27346623	RNU6ATAC	6.37E-06
A_17_P16966417	TIAF1	3.99E-05
A_17_P09879614	TM2D3	4.71E-05
A_17_P08379544	CTDSP2	5.27E-05
A_17_P16292171	DHTKD1	6.24E-05
A_17_P17194852	DBNDD2	6.48E-05
A_17_P04327834	PCDHGC4	9.10E-05

The Brain-All subset revealed the largest number of relevant enriched biological pathways. To confirm that the identified biomarkers capture signal and not noise, we conducted 100 permutations of sample labels. At each permutation, the enriched pathways were obtained from KEGG and Biocarta, using WebGestalt. None of the pathways in Table 3.8 were significant during permutations. This verifies that the obtained biomarkers are biologically reliable. Table 3.8 lists the enriched biological pathways using the biomarkers from Brain-All (Table 3.5) using the fuzzy rule-based approach.

Table 3.8 List of enriched pathways using the genes in Table 3.5. Neurologically relevant pathways are highlighted in green and bold.

p-value	**Pathway Name and Description**
7.63E-08	**Alzheimer's disease.**
2.04E-06	**Long-term Potentiation (LTP):** a persistent strengthening of synapses based on recent patterns of activity. These are pathways of synaptic activity that produce a long-lasting increase in signal transmission between two neurons. LTP is widely considered one of the major cellular mechanisms that underlies learning and memory.
2.84E-06	VEGF signaling pathway: Vascular Endothelial Growth Factors stimulate vascular endothelial cell growth, survival, and proliferation.
1.07E-05	Amoebiasis: an infection of the colon.
1.15E-05	T Cell Receptor signaling pathway: TCR activation promotes a number of signaling cascades that determine cell fate.
2.31E-05	**Axon guidance signaling pathway:** a key stage in the formation of neuronal networks.
2.41E-05	**MAPK signaling pathway:** a chain of proteins in the cell that communicates a signal from a receptor on the surface of the cell to the DNA in the nucleus.
4.18E-05	**WNT signaling pathway:** involved in a wide range of cellular activities. For example, Wnt1 antagonizes neural differentiation and is a major factor in self-renewal of neural stem cells. This allows for regeneration of nervous system cells, indicating a role in promoting neural stem cell proliferation
0.0001	Focal adhesion: focal adhesions are the sub-cellular structures that mediate the regulatory effects (i.e., signaling events)
0.0001	Adherens junction: protein complexes that occur at cell–cell junctions in epithelial and endothelial tissues.
0.0001	**B cell receptor signaling pathway:** The complexity of BCR signaling permits many distinct outcomes, including survival, tolerance (anergy) or apoptosis, proliferation, and differentiation into antibody-producing cells or memory B cells.
0.0002	Apoptosis: is a process of programmed cell death that occurs in multicellular organisms.
0.0003	Melanogenesis: related to the color of eye and skin etc.
0.0003	GnRH signaling pathway: is a key regulator of the reproductive system, triggering the synthesis and release of LH and FSH in the pituitary.
0.0004	Oocyte meiosis: is the creation of an ovum (egg cell).
0.0006	Osteoclast differentiation: related to bone structure.
0.0021	**Amyotrophic lateral sclerosis (ALS)**: a progressive neurodegenerative disease that affects nerve cells in the brain and the spinal cord.
0.0026	Regulation of actin cytoskeleton: can lead to diverse effects on cell activity, including changes in cell shape, migration, proliferation, and survival.
0.0036	**Long-term depression**
0.0115	**Neurotrophin signaling pathway:** Neurotrophins are a family of trophic factors involved in differentiation and survival of neural cells. The neurotrophin family consists of nerve growth factor (NGF), brain derived neurotrophic factor (BDNF), neurotrophin 3 (NT-3), and neurotrophin 4 (NT-4).

3.2 Application of COMBINER to PTSD datasets

Application of COMBINER to PTSD DNA Methylation dataset using Allen Brain Atlas and Promoter Region subsets

COre Module Biomarker Identification with Network Exploration (COMBINER) is a feature selection tool, developed by Yang et al. [48], which takes into account the variability in genomic data across different tissue and subjects. Using multiple cohorts of data as input, it enables the identification of disease genes, pathway biomarkers and the construction of their associated regulatory networks. In our analysis, we apply COMBINER to the PTSD DNA methylation dataset (Agilent platform), consisting of measurements of 237,117 probes collected from PTSD (50 subjects) and control (51 subjects) groups. After removing probes that do not correspond to known human genes, we are left with 183,310 probes in total.

We use the extended version of COMBINER [49], which enables us to build classifiers for diagnostic applications. In this version, the given dataset is partitioned into three cohorts with equal numbers of subjects from the two groups. The first cohort is used to infer the core modules for every pathway. The top 100 features with most discriminative power are then used for downstream analysis. The data from the second cohort is projected onto the top 100 pathways. The Consensus Feature Elimination (CFE) algorithm is run 20 times, and a voting strategy is employed to identify the candidate biomarker pathways. Using the third cohort we compute the cross-validated training performance on the identified biomarkers. The whole dataset is partitioned 20 times and the steps enumerated above are repeated on all partitions. The final set of predicted biomarker pathways were obtained by a voting strategy over the set of all candidate pathways discovered in different partitions. The results obtained by COMBINER on the four previously described data subsets (Brain-All, Brain-Promoter, All, and Promoter Only) are shown in Figure 3.3 and Table 3.9.

In order to obtain a comprehensive biological picture, we performed the pathway enrichment analysis using a hypergeometric test. This was done on the best performing panel, which was obtained by considering only the promoter region of the genes found in blood. The network of pathways is shown in Figure 3.4.

Table 3.9 Table of classification performance. The "predicted" performance from cross-validation (P) and "actual" performance based on independent validation (A) for the four analyzed datasets are show. For each dataset, the number of candidate biomarker probes, percentage of probes with p<0.001 and percentage of probes in the promoter region of the gene are also shown.

AUC: area under Receiver Operating Characteristic (ROC) curve; ER: Error Rate; TPR: True Positive Rate; FPR: False Positive Rate

Metric	Promoter Only		Brain-All		Brain-Promoter		All	
	P	A	P	A	P	A	P	A
AUC	0.7642 ± 0.1103	0.6658	0.6286 ± 0.1105	0.4921	0.6857 ± 0.0646	0.5500	0.7960 ± 0.0793	0.5579
ER	0.3047 ± 0.0872	0.3846	0.3981 ± 0.0921	0.5128	0.3703 ± 0.0593	0.4359	0.2859 ± 0.0797	0.5385
TPR	0.6921 ± 0.0629	0.6500	0.5970 ± 0.1118	0.5000	0.6379 ± 0.0860	0.6500	0.7362 ± 0.1148	0.6000
FPR	0.3015 ± 0.1361	0.4311	0.3932 ± 0.1342	0.5263	0.3785 ± 0.0679	0.5263	0.3081 ± 0.0843	0.6842
No. of probes	153		24		84		65	
% probes with p<0.001	60.78%		50.00%		41.66%		89.23%	
% promoter probes	100.00%		25.00%		100.00%		30.77%	

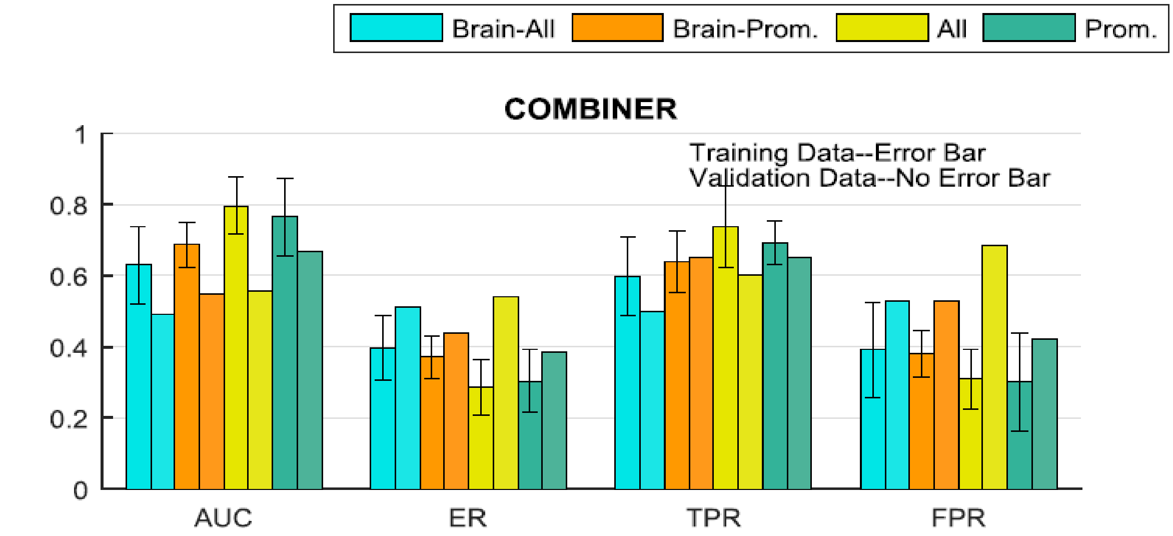


Figure 3.3 Performance metric results obtained by COMBINER. The bars with error bars indicate the mean and standard deviation of four performance metrics from cross-validation in training data. The height of bars without error bars indicates performance in validation dataset.

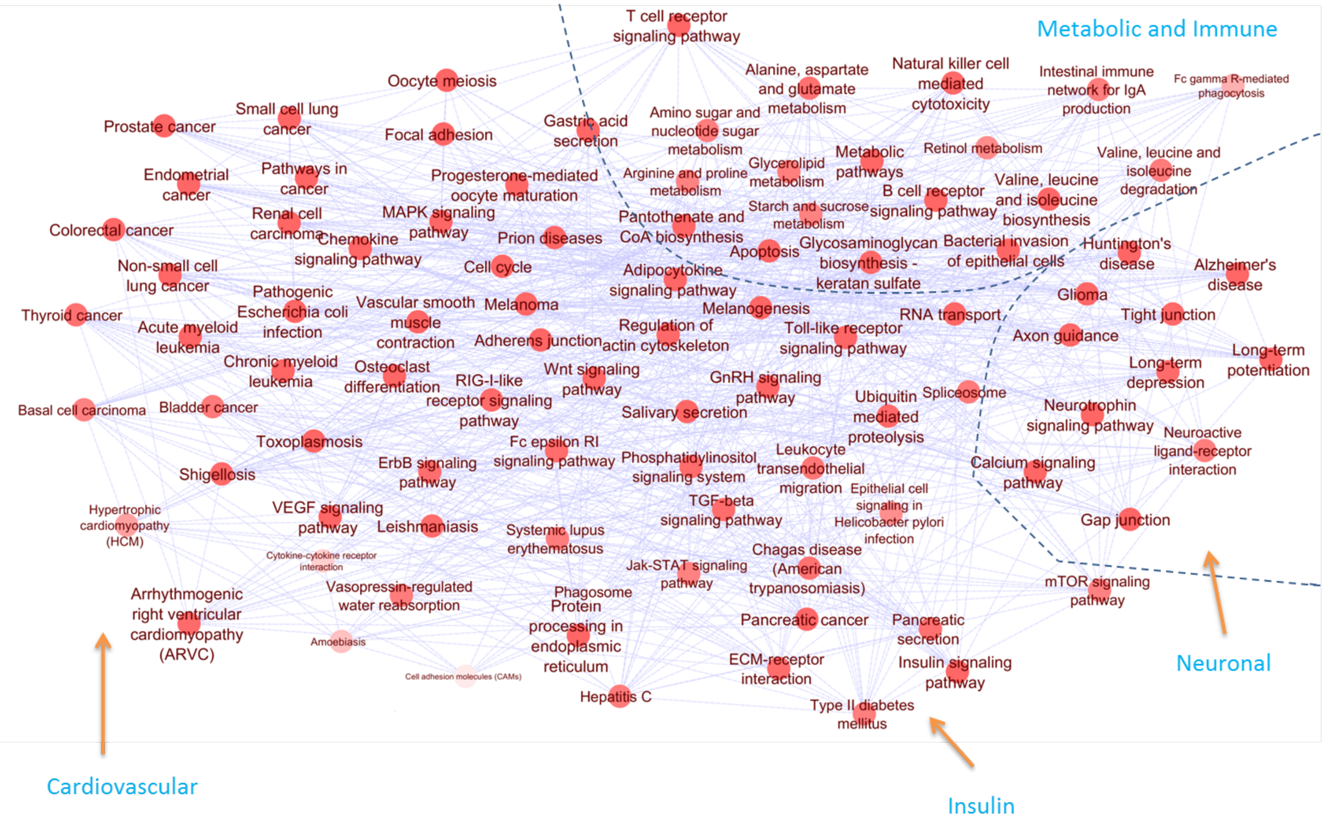

Metabolic and Immune
Neuronal
Cardiovascular
Insulin
T cell receptor signaling pathway
Oocyte meiosis
Alanine, aspartate and glutamate metabolism
Natural killer cell mediated cytotoxicity
Intestinal immune network for IgA production
Fc gamma R-mediated phagocytosis
Small cell lung cancer
Prostate cancer
Focal adhesion
Gastric acid secretion
Amino sugar and nucleotide sugar metabolism
Retinol metabolism
Valine, leucine and isoleucine degradation
Endometrial cancer
Pathways in cancer
Progesterone-mediated oocyte maturation
Arginine and proline metabolism
Glycerolipid metabolism
Metabolic pathways
B cell receptor signaling pathway
Valine, leucine and isoleucine biosynthesis
Colorectal cancer
Renal cell carcinoma
MAPK signaling pathway
Pantothenate and CoA biosynthesis
Starch and sucrose metabolism
Chemokine signaling pathway
Prion diseases
Apoptosis
Glycosaminoglycan biosynthesis - keratan sulfate
Bacterial invasion of epithelial cells
Huntington's disease
Non-small cell lung cancer
Cell cycle
Adipocytokine signaling pathway
Alzheimer's disease
Pathogenic Escherichia coli infection
Vascular smooth muscle contraction
Melanoma
Melanogenesis
Glioma
Thyroid cancer
Acute myeloid leukemia
Regulation of actin cytoskeleton
RNA transport
Tight junction
Toll-like receptor signaling pathway
Axon guidance
Adherens junction
Long-term potentiation
Chronic myeloid leukemia
Osteoclast differentiation
RIG-I-like receptor signaling pathway
Wnt signaling pathway
GnRH signaling pathway
Long-term depression
Ubiquitin mediated proteolysis
Spliceosome
Basal cell carcinoma
Bladder cancer
Salivary secretion
Neurotrophin signaling pathway
Neuroactive ligand-receptor interaction
Toxoplasmosis
Fc epsilon RI signaling pathway
Phosphatidylinositol signaling system
Leukocyte transendothelial migration
Shigellosis
ErbB signaling pathway
Calcium signaling pathway
Epithelial cell signaling in Helicobacter pylori infection
Hypertrophic cardiomyopathy (HCM)
VEGF signaling pathway
TGF-beta signaling pathway
Leishmaniasis
Systemic lupus erythematosus
Gap junction
Cytokine-cytokine receptor interaction
Chagas disease (American trypanosomiasis)
Jak-STAT signaling pathway
mTOR signaling pathway
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Vasopressin-regulated water reabsorption
Phagosome
Protein processing in endoplasmic reticulum
Pancreatic cancer
Pancreatic secretion
Amoebiasis
ECM-receptor interaction
Insulin signaling pathway
Cell adhesion molecules (CAMs)
Hepatitis C
Type II diabetes mellitus

Figure 3.4 Network of pathways obtained by the enrichment analysis for the candidate biomarker probes found when considering only the promoter region of all genes. Two pathways are connected in the network if they share a common candidate biomarker gene. Pathways corresponding to similar biological processes such as metabolic and immune function, or neuronal proceses are grouped together.

Application of COMBINER to other single ‘omic PTSD datasets

We completed single omics analysis on the 52-52 discovery cohort from five different modalities. These included methylation (Illumina platform), mRNA, protein, microRNA and metabolite datasets. The multi-metric performance results for these single omics datasets are shown in Table 3.10.

Table 3.10 Performance metric results obtained by COMBINER for different single omic datasets. We analysised five different modalities each having at most 51 subject with PTSD and 51 subject from the control group. Validation was performed on an independent dataset comprising of 31/31 subjects. P: Predictive cross-validated performance on the 51/51 discovery set, A: Actual or test performance on the validation dataset. The protein data comprised of a very small subet of subjects, hence no predictive cross-validated performance was computed.

	DNA Methylation (Illumina)		**mRNA**		**miRNA**		**Protein**		**Metabolite**	
	P	A	P	A	P	A	P	A	P	A
AUC	0.508±0.133	0.514	0.507±0.163	0.668	0.613±0.089	0.618	-	0.646	0.613±0.120	0.668
ER	0.491±0.113	0.436	0.482±0.124	0.328	0.394±0.075	0.421	-	0.367	0.413±0.095	0.323
TPR	0.490±0.128	0.548	0.509±0.138	0.714	0.517±0.086	0.548	-	0.548	0.553±0.093	0.613
FPR	0.472±0.116	0.419	0.475±0.131	0.367	0.305±0.130	0.385	-	0.276	0.379±0.118	0.258
# of biomarkers	2		20		3		9		16	

3.3 Selective reaction monitoring (SRM) quantitative proteomic approach

We are conducting comprehensive proteomics analyses to identify blood proteins that can be used as biomarkers for PTSD diagnosis. SRM is the main approach that we are using to identify blood protein biomarkers for PTSD. It is a sensitive protein quantitation method based on a two-stage mass filtering in a triple quadruple (QqQ) mass spectrometer (MS). This targeted proteomics approach enhances the detection limit (as low as fmol level) of selected proteins in complex biological samples. SRM provides the capacity to measure as many as 100-150 proteins in a single 2-hour MS run [50].

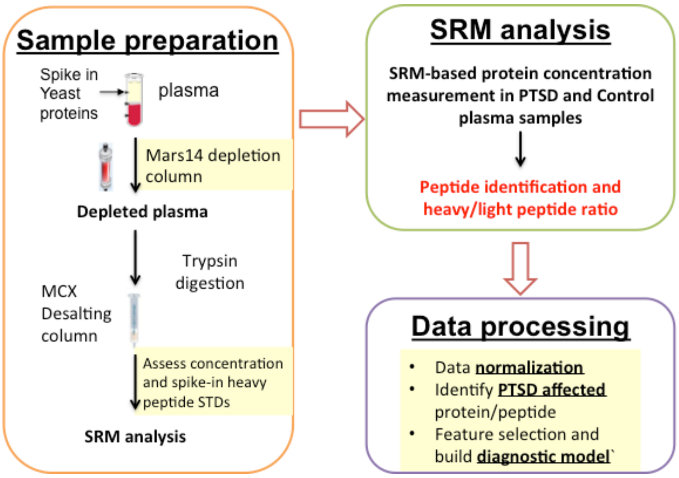

Sample preparation
Spike in Yeast proteins
plasma
Mars14 depletion column
Depleted plasma
Trypsin digestion
MCX Desalting column
Assess concentration and spike-in heavy peptide STDs
SRM analysis
SRM analysis
SRM-based protein concentration measurement in PTSD and Control plasma samples
Peptide identification and heavy/light peptide ratio
Data processing
Data normalization
Identify PTSD affected protein/peptide
Feature selection and build diagnostic model`

Figure 3.5 Workflow for circulating protein biomarker discovery using targeted proteomics.

Since SRM is a targeted proteomics approach, selecting proper proteins is critical to achieve fruitful results. To obtain a comprehensive list of protein candidates, we used five different but complementary approaches as follows: i) proteins that are preferentially expressed in major organs in the body (brain, heart, liver, lung, kidney, and white blood cells); ii) proteins that were identified through global profiling studies with plasma samples from humans (normal controls and soldiers with PTSD) and mice (a social defeat mouse PTSD model); iii) proteins that were differentially expressed in different brain subregions in a PTSD mouse model; iv) PTSD, TBI, and anxiety disorder-associated proteins identified in literature; and v) biomarker candidates identified within the consortium (such as BDNF, GRIA1, CLOCK, TGFB1). In total, we assembled a list of 1,043 proteins for the initial screen (**Table 3.11**). These proteins reflect a number of important biological pathways and processes associated with neuropathophysiological functions such as neuroactive ligand-receptor interaction, tyrosine metabolism, and long-term potentiation.

At the initial screening stage, all 1,043 proteins were monitored from both 52/52 and 31/31 sample sets. The overall SRM workflow is shown in Figure 3.5. The number of proteins in each category that were measured and detected is summarized in Table 3.11. The SRM data were first analyzed with the Skyline program. The Light/Heavy peptide ratios, which reflect the abundance of target peptides (proteins), were then extracted. In total, we can reliably detect and quantify 89 proteins in the plasma by SRM. As expected, more than half of these detectable proteins are from the liver (53 out of 89, Table 3.10) since the liver contributes to a significant portion of blood proteins.

Table 3.11 Summary of selected proteins in each category for SRM based measurement.

Group		Number of Proteins Screened	Number of Proteins Detected
Enriched in specific organ	Brain	202	15
	Heart	78	3
	Kidney	47	0
	Liver	165	53
	Lung	30	4
	WBC	33	8
PTSD associated protein	Identified from iTRAQ and mouse model	305	8
	Identified from literature	186	9
	Identified within the consortium	23	2
Total#		1043	89

***#:** after remove duplicated proteins from different lists*

The concentration of these 89 proteins was then measured in the 83/83 cohort set by SRM. All 89 proteins (150 peptides) can be measured in a single SRM run. After normalizing the data based on original plasma sample volume, peptide abundances from the 83/83 sample sets were assembled for further analysis.

Correlations between SRM protein measurement results and clinical information

To examine whether patient clinical information is associated with any of the plasma protein concentrations measured by SRM assays, we computed the Pearson correlation coefficient between clinical information and peptide (protein) measurement results. This comparative analysis showed that the level of CRP (C-reactive protein) measured by SRM correlates well with concentrations determined by ELISA in clinical lab (**Figure 3.6**). The analysis also showed a good correlation between plasma Apolipoprotein concentrations and blood triglyceride levels, reflecting the involvement of the APOE protein at the triglyceride level. These results suggest the accuracy and reproducibility of the SRM-based protein concentration measurement.

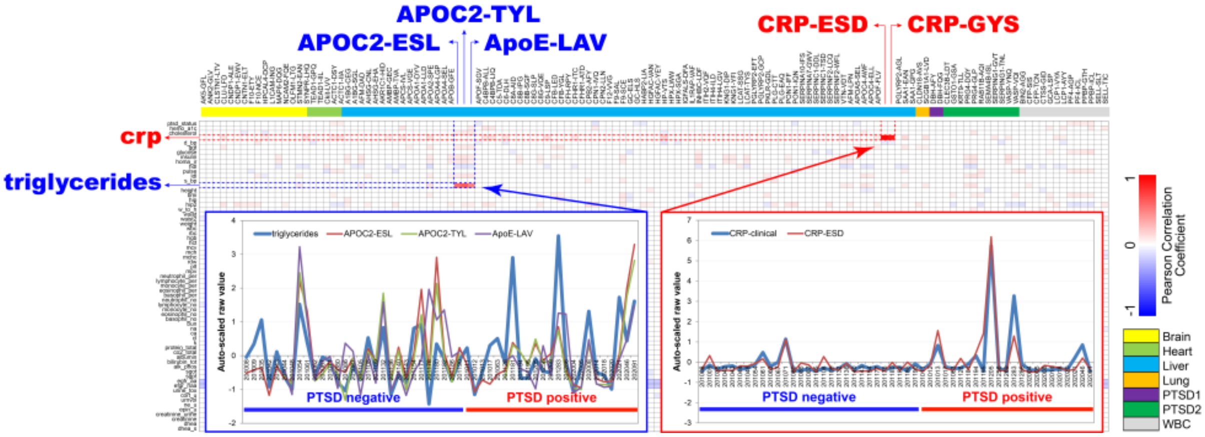


Figure 3.6 Correlations between plasma proteins (peptides) (APOC2, APOE and CRP) and clinical measures (crp and triglycerides).

<u>Identification of blood protein biomarker panel by Selective Reaction Monitoring (SRM) proteomic approach</u>

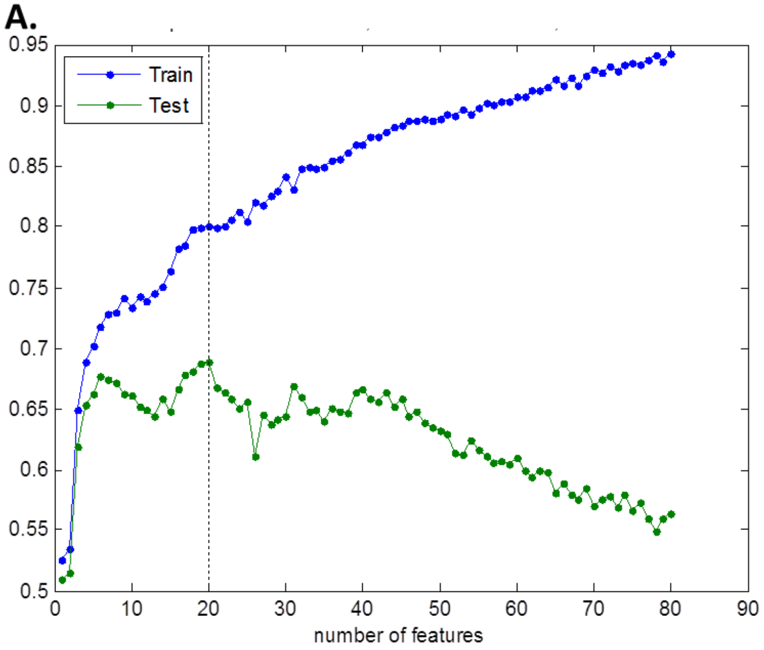
We received additional plasma samples from the original 52/52 discovery and 31/31 validation sample sets last spring so that we can complete the SRM and miRNA analyses on the entire set of samples. With the newly acquired samples, we conducted SRM and miRNA analyses for the entire 83/83 sample sets. The plasma samples were analyzed with SRM in duplicates to test the reproducibility of the measurement. The summary of the SRM sample composition for the 83/83 set is listed in Table 3.12.

Table 3.12 SRM sample composition from the combined sets of 83/83 (52/52 and 31/31).

	52/52		31/31	
	PTSD-	**PTSD+**	**PTSD-**	**PTSD+**
201	43	15	18	13
202	8	36	11	18
Total	**52**	**51**	**29**	**31**

The SRM raw data were processed with Skyline 3.5.0 to quantify peak areas of fragment ions from target peptides. The intensities of peptides were normalized by the corresponding heavy isotope labeled spiked-in peptides. The ratios of the intensities between endogenous light peptides and heavy peptides were calculated and transformed to log with base 2. In total, we quantified 96 peptides representing 69 proteins from 162 samples of the 83/83 sample set.

Based on all the data obtained, we observed some batch effects in the three groups of samples. However, the batch difference was reasonably reduced when applying a batch effect correction method. The SRM data were then analyzed and a panel of top ranked 20 peptides (translated into 18 unique proteins) gave good average classification performances based on 100-times 5-fold cross validation using a support vector machine in the 83/83 samples. The top ranked 20 peptides are listed in Figure 3.7B. The combined 20-peptide panel provided an AUC of 0.756 with an accuracy of 68.8% in the discovery dataset (**Figure 3.7A**).



B.

20 selected peptides:

CPN1-IVQ
HPX-SGA
C4BPB-LIQ
CTSS-GID
CLU-LFD
PGLYRP2-GCP
VTN-VDT
APOE-LGA
APCS-IVL
B7Z8Q5-LIA
ACTC1-DSY
AMBP-GEC
AFM-LPN
APOF-SGV
ITIH2-VQF
C5-TDA
IL1RAP-VAF
ITIH4-LGV
PLG-EAQ
PGLYRP2-EFT

Figure 3.7. Summary of 5-fold Cross validation performance on peptide panels. The performance curve showed that the 20 peptides panel provides an optimal feature set (AUC: 0.756 and ACC: 68.8%). AUC: Area under the curve, ACC: Accuracy.

3.4 Identification of miRNA biomarker panels based on small RNA sequencing data

A more reliable small RNA library construction method

Next generation sequencing (NGS) is gaining interest as the method of choice to profile small RNA including miRNA in biological samples. It is well-known that current small RNA library construction methods suffer from significant sequence bias [51-54]. Depending on which library construction kit is used, very different profiling results can be obtained from the same sample. For example, hsa-miR-486 is the most abundant miRNA identified in libraries constructed with the Illumina TruSeq kit (accounting for > 80% of the reads in some libraries), whereas with the NEB (New England Biolabs) kit it typically accounts for only about 5%. This creates significant downstream validation problems.

We developed a small-RNA library construction procedure that offers higher reproducibility with less sequence bias (**Figure 3.8**). In the new protocol, we have incorporated four degenerate bases in both the 3-prime and 5-prime adapters which provide a more favorable ligation partner for each individual miRNA thus reducing ligation bias compared to traditional approaches. In addition, we replaced the typical RNA ligase used in most commercial kits with a modified version that produces fewer side products. To enhance ligation efficiency, we also included polyethylene glycol (PEG) in the ligation mixture. We used a single-stranded DNA binding protein, a 5' deadenylase enzyme, and a DNA specific exonuclease to degrade any unligated RNA adapters, which also suppress the formation of unwanted ligation products such as adapter dimers in the library. To further reduce the amplification of any remaining adapter-dimer, we adapted a two-stage size-selection step in which an initial size selection is performed after 4 cycles of amplification. With this new library construction method, we have significantly increased the number of reads mapped to miRNA and reduced the bias in both synthetic miRNA as well as real biological sample.

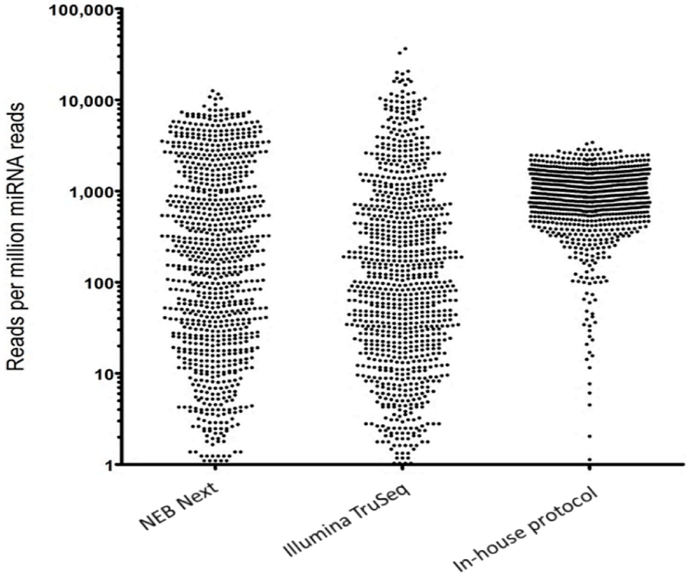


Figure 3.8. New small RNA library construction protocol reduces sequence bias. The read distribution in sequence libraries generated from a pool of 962 equal molar synthetic miRNAs. Y-axis: read per million, X-axis: different library construction method.

MiRNA profiles of all 83/83 set plasma samples were generated with small RNA sequencing technology except two samples (**Table 3.13**). The samples were run in four different sequencing runs due to the number of samples can be on the sequencing flow cell.

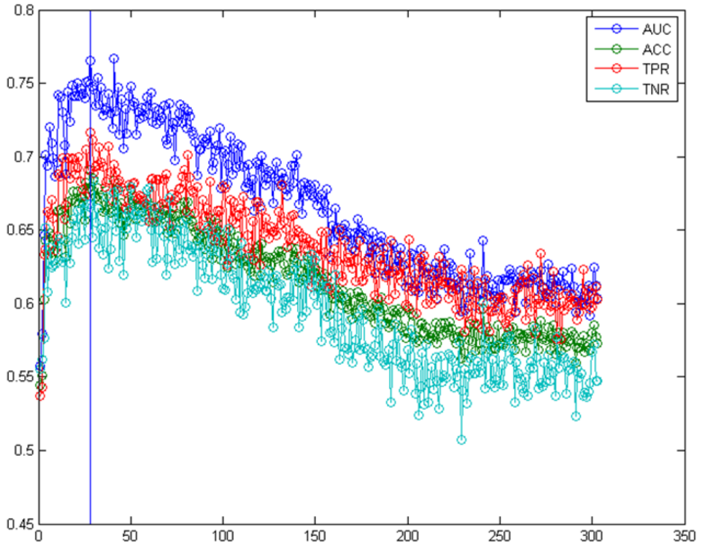
Table 3.13 Information of samples for small RNA sequencing data.

Batch	PTSD Negative	PTSD Positive	Total Sample
Batch 1.	12	24	36
201 samples	12	12	24
202 samples	0	12	12
Batch 2.	43	5	48
201 samples	40	3	43
202 samples	3	2	5
Batch 3.	12	36	48
201 samples	2	10	12
202 samples	10	26	36
Batch 4.	15	17	32
201 samples	7	3	10
202 samples	8	14	22
Total	**82**	**82**	**164**

We used an in-house developed small RNA library construction method – 4N protocol to reduce miRNA measurement bias in the current commercial library construction kits. The 4N protocol significantly reduced library construction bias as indicated in Figure 3.8 when a pool of equal molar of synthetic miRNAs were used to construct small RNA library.

Classification analysis of sequencing data

Similar to the approach used on the SRM data, the SVM-RFE feature selection algorithm was applied to small RNA sequencing data to identify optimal feature set showing maximum classification performance. With 100 times 5-fold cross validation, we identified 28 miRNAs panel provides an AUC of 0.7655 (Accuracy: 0.6906, Sensitivity: 0.7167, Specificity: 0.6674) (**Figure 3.9**).



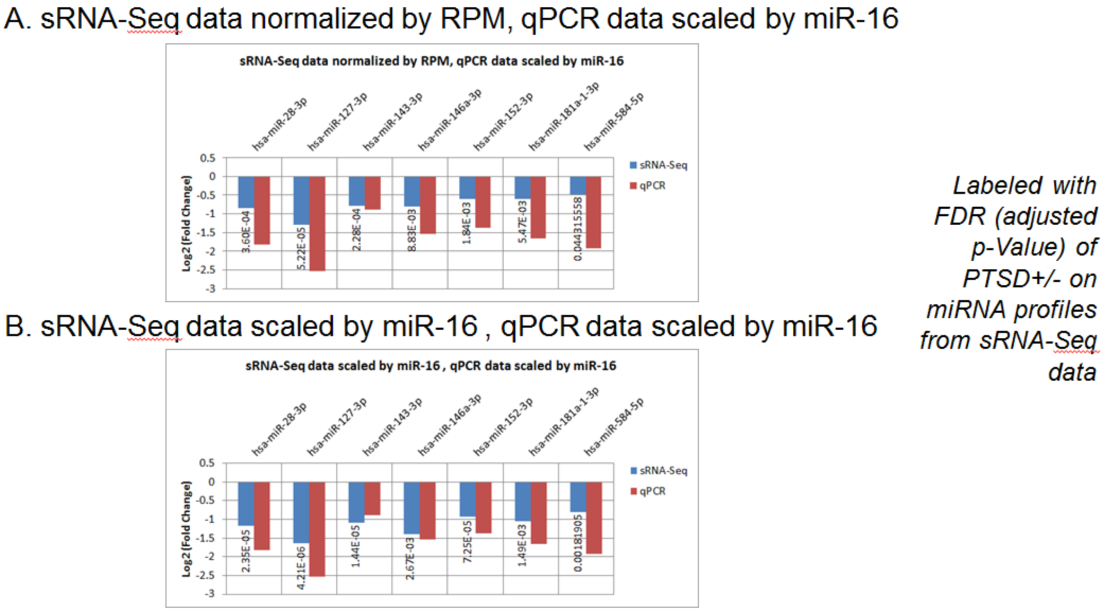
28 selected miRNA:

hsa-miR-598-3p
hsa-miR-4286-5p
hsa-let-7b-5p
hsa-miR-590-3p
hsa-miR-192-5p
hsa-miR-99a-5p
hsa-miR-155-5p
hsa-miR-100-5p
hsa-miR-9-1-5p
hsa-miR-363-3p
hsa-miR-133a-2-3p
hsa-miR-484-5p
hsa-miR-144-3p
hsa-miR-182-5p
hsa-miR-4485-3p
hsa-miR-3615-3p
hsa-miR-4317-5p
hsa-miR-93-3p
hsa-miR-133a-1-3p
hsa-miR-338-3p
hsa-let-7e-5p
hsa-miR-93-5p
hsa-miR-329-2-3p
hsa-miR-660-5p
hsa-miR-4454-5p
hsa-miR-4532-5p
hsa-miR-505-3p
hsa-miR-548au-5p

Figure 3.9 Summary of 5-fold Cross validation performance on miRNA panels. The performance curve showed that the 28 miRNAs panel provides an optimal feature set, indicated by vertical blue line (AUC: 0.7655 and ACC: 69.06%). AUC: Area under the curve, ACC: Accuracy, TPR: True positive rate (Sensitivity), TNR: True negative rate (Specificity).

Validation of miRNA-sequencing data by RT-qPCR

To validate the results from the small RNA sequencing data, we conducted qPCR with advanced Taqman miRNA assay on 7 selected human miRNAs: miR-28-3p, miR-127-3p, miR-143-3p, miR146a-3p, miR-152-3p, miR-181a-1-3p and miR-584-5p. Several invariant miRNAs were also included and used in qPCR normalization. Due to the limitation of amount of the plasma samples we have, only 128 samples (70 control and 58 PTSD patients) were used. To check the consistency of qPCR data, three replicates were generated for each of the miRNAs. Standard deviation of the three replicates was then computed and sample with standard deviation value greater than 3 were removed from analysis. Comparing results between qPCR and sequencing data showed a good correlation, which suggest the sequencing based miRNA profiling results are comparable with qPCR. The fold changes of these 7 miRNAs between PTSD positive and negative patients observed in sequencing can also be validated in qPCR data (**Figure 3.10**).



A. sRNA-Seq data normalized by RPM, qPCR data scaled by miR-16

B. sRNA-Seq data scaled by miR-16 , qPCR data scaled by miR-16

Labeled with FDR (adjusted p-Value) of PTSD+/- on miRNA profiles from sRNA-Seq data

Figure 3.10 Comparison of fold changes (PTSD+/-) between small RNA sequencing data (blue bars) and qPCR data (red bars).

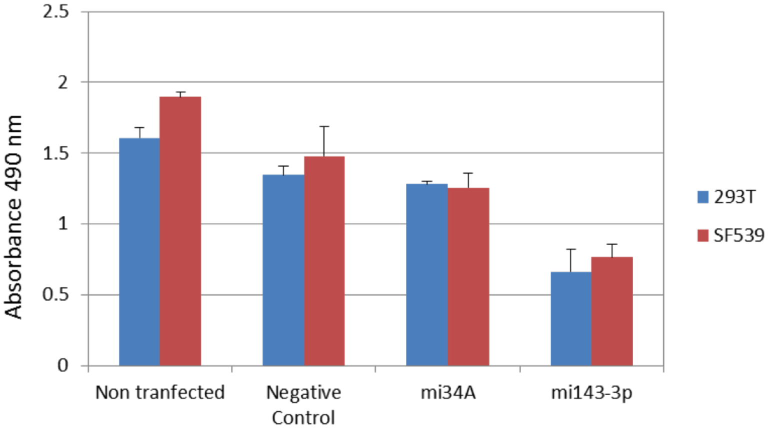


Figure 3.11 In vitro cell proliferation was determined by MTS based assay after 48 hours post transfection with miR-143-3p mimic and controls.

3.5 Functional implication of PTSD associated circulating miRNA

Our results showed a number of dysregulated circulating miRNAs in PTSD patients. Among these, the concentration of a meninges enriched miRNA, miR-143-3p, is decreased in PTSD patients. We are interested in further understanding the role of miR-143-3p in PTSD. Interestingly, some of the predicted targets of miR-143-3p such as NTRK2, BCL2 , GABRB3, CALM1, EGR1 and TSC22D3, have been shown to play a role in PTSD [55-58]. To assess the function of miR-143-3p, we transfect the microRNA mimics into glioblastoma cell line (SF539) and 293T (epithelial cells). The overexpression of miR-143-3p inhibits/reduces the proliferation of these cells based MTS assays (**Figure 3.11**). Additionally, immunofluorescence staining of Col1A1, a validated target of miR-143-3p, suppresses the level of the Col1A1 protein in SF539 cells (**Figure 3.11**).

3.6 Integration of SRM and miRNA data for biomarker panel discovery

We applied support vector machine (SVM) with recursive feature elimination (SVM-RFE) algorithm to find an optimal subset of features to classify PTSD- and PTSD+ groups. SVM-RFE was applied to select peptides and miRNAs for the classification of PTSD- and PTSD+ groups. The 5-fold cross-validation was repeated 100 times to identify optimized features and obtain an unbiased estimation of classification accuracy. The importance of each feature in the classification was determined based on the selection frequency from 5-fold cross-validations. The features were sorted in order of their frequencies. By increasing the number of features, SVM models were constructed and the average classification accuracies were computed. The optimal feature set was then determined at the highest average classification accuracy of the test set. By applying repeated cross validations with SVM-RFE to 83/83 sample set, we identified 20 peptides with AUC of 0.756 in cross validation test sets. The performance of the identified 20 peptides was validated in 33 validation samples. The AUC, accuracy, sensitivity and specificity were 0.7545, 0.6061, 0.5769, and 0.6250, respectively.

Then SRM and miRNA-seq data were also integrated by combining 20 peptides and 28 miRNAs identified from the analyses using individual data. The same algorithm was applied and 24 features of 15 peptides and 9 miRNAs were identified with AUC, accuracy, sensitivity, and specificity of 0.8305, 0.7784, 0.7488, and 0.8081, respectively. The performances using peptides, miRNAs, and both datasets are summarized in Table 3.14.

Table 3.14 Summary of classification performance. The performance is based on 5-fold cross validation using 83/83 set.

Data type (Data set)	**Peptide (83-83)**	**miRNA (83-83)**	**Integration (83-83)**
No of feature(s)	**20**	**28**	**24**
Feature(s)	CPN1-IVQ, HPX-SGA, C4BPB-LIQ, CTSS-GID, CLU-LFD, PGLYRP2-GCP, VTN-VDT, APOE-LGA, APCS-IVL, B7Z8Q5-LIA, ACTC1-DSY, AMBP-GEC, AFM-LPN, APOF-SGV, ITIH2-VQF, C5-TDA, IL1RAP-VAF, ITIH4-LGV, PLG-EAQ, PGLYRP2-EFT	miR-598-3p, miR-4286-5p, let-7b-5p, miR-590-3p, miR-192-5p, miR-99a-5p, miR-155-5p, miR-100-5p, miR-9-1-5p, miR-363-3p, miR-133a-2-3p, miR-484-5p, miR-144-3p, miR-182-5p, miR-4485-3p, miR-3615-3p, miR-4317-5p, miR-93-3p, miR-133a-1-3p, miR-338-3p, let-7e-5p, miR-93-5p, miR-329-2-3p, miR-660-5p, miR-4454-5p, miR-4532-5p, miR-505-3p, miR-548au-5p	**15 peptides** : C4BPB-LIQ, CPN1-IVQ, HPX-SGA, PGLYRP2-GCP, CLU-LFD, B7Z8Q5-LIA, APCS-IVL, CTSS-GID, PLG-EAQ, AFM-LPN, ITIH2-VQF ,PGLYRP2-EFT, VTN-VDT, C5-TDA, ACTC1-DSY **9 miRNAs** : miR-4532-5p, let-7b-5p, miR-100-5p, miR-338-3p, miR-598-3p, miR-93-5p, miR-155-5p, miR-4485-3p, miR-4317-5p
AUC	0.756	0.7655	0.8305
ACC	0.688	0.6906	0.7784
TPR	0.681	0.7167	0.7488
TNR	0.691	0.6674	0.8081

3.7 Development of optimization strategies to improve sensitivity or specificity for biomarker identification and classification

Metric-focused recursive feature elimination for customized biomarker identification

Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. The estimator ranks features by estimating the impact of projecting subjects from an N-dimensional space to a smaller (N-1 dimensional) subspace. Most of these approaches are insensitive to the misclassification error rate of a given class. We have developed a new approach in which the estimator takes into account the misclassification rate of each class. The RFE approach is used for feature reduction to obtain biomarkers, in conjunction with a classifier. In this work, we chose to focus on linear classifiers. The benefit of the proposed method has been demonstrated for a binary classification problem on a highly replicated publicly available TCGA ovarian cancer dataset.

In the present work, we use linear discriminant analysis (LDA) and support vector machine (SVM) to obtain the linear decision boundary. In the traditional RFE method (baseline), the decision to remove a feature is based on the coefficients of the decision boundary obtained by a linear classifier. Features are sequentially removed and the subset of features which gives the best performance is considered as the final feature subspace. This greedy search approach does not guarantee a global optimal solution. The proposed method modifies the feature elimination step and takes into account the class-specific classification errors. The analysis pipeline consisting of bootstrapping, cross-validation and recursive feature elimination is shown in Figure 3.12.

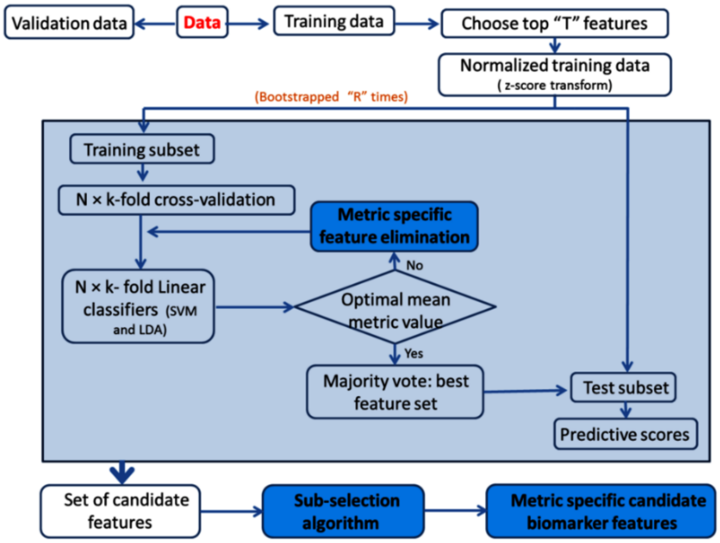


Figure 3.12 Data analysis pipeline for metric-specific biomarker identification. Bootstrapping and k-fold cross validation are used to obtain robust features. The metric-specific feature elimination algorithm is implemented during cross-validation to get the best feature set.

Metric-specific feature elimination method

Any hyperplane can be written as

$$w_i \cdot x_i + b = 0. \tag{1}$$

We can rewrite Eqn. 1 as

$$\frac{w_i \cdot x_i}{\sqrt{\sum w_i^2}} + \frac{b}{\sqrt{\sum w_i^2}} = 0 \tag{2}$$

The above equation can be written in the Hessian normal form

$$\boldsymbol{n} \cdot \boldsymbol{x} = -p, \tag{3}$$

where $n_i = \frac{w_i}{\sqrt{\sum w_i^2}}$ are the direction cosine of the unit vector perpendicular to the given plane. The RFE strategy proposed by [59] inherently uses the values of the direction cosines of the decision boundary obtained by support vector machine to rank features. In this top down feature reduction approach, the feature with the smallest coefficient in Eqn. 1 is removed. This is equivalent to removing a feature with the smallest direction cosine value (the largest angle). Principal Component Analysis (PCA) is a widely used tool to extract statistical information from a given dataset. It employs the covariance structure of the data to generate an orthogonal vector set referred to as principal components. Each of these vectors is the eigenvector of the covariance matrix and the associated eigenvalue is the measure of relative variance in that direction.

In the proposed method we modify the RFE ranking strategy or feature importance estimator function, to incorporate class specific variance information. This allows for the selection of features which have expression values c_i more tightly regulated in one group vs another. We incorporate this information into the feature importance estimator function by considering the first principal components of only the class c_i. More precisely, we define

$$\widetilde{f_i} = \frac{w_i}{u_i} \tag{4}$$

Where u_i are the direction cosine of the first eigenvector of the subset of data that belongs to class c_i. Every coefficient w_i of the decision boundary (DB) can be imagined to be weighted by the direction cosine of the eigenvector of the class c_i. The feature that minimizes $\tilde{f}$ is removed. The feature removed then enables the projection of the data onto a subspace that is closest to the orthogonal subspace to the DB and simultaneous reduces the variance of c_i. Using the above method, if there exists a set of highly discriminative features such that there are significant differences in relative variances for a given class, then one can obtain lower misclassification rate for that class. For example, reducing to relative variance of one specific class may lead to improvement in the misclassifications of subjects in that class.

Application of metric-specific feature elimination to ovarian cancer dataset

An overview of the metric-specific biomarker identification is shown in Figure 3.12. Using ovarian cancer Affymetrix gene expression data from The Cancer Genome Atlas (TCGA), biomarker identification was performed to classify subjects based on mortality. A total of 159 positive subjects (survival less than three years), and 160 negative subjects (survival more than three years) were used. The biomarker performance metrics for cross-validation and validation are shown in Figure 3.13. Similar performance is achieved for the ER, TPR and FPR-estimator strategy, with slightly improved TPR performance using the TPR-estimator, and a lower False Positive Rate using the FPR-estimator strategy.

The distribution of gene expression of the final biomarker candidates for each strategy is shown in Figure 3.14. As expected, the TPR-estimator strategy identifies biomarkers with greater expression variance in the negative class, while the FPR-estimator strategy identifies biomarkers with greater expression variance in the positive class. Next, we compared the candidate biomarker panels from each strategy, computing the overlaps between panels and the fraction of candidate biomarkers genes that are known to be cancerous from NetPath [60], Atlas of Cancer Genes [61], Census Genes [62], G2BC [63], and KEGG Pathways of Cancer [64]).

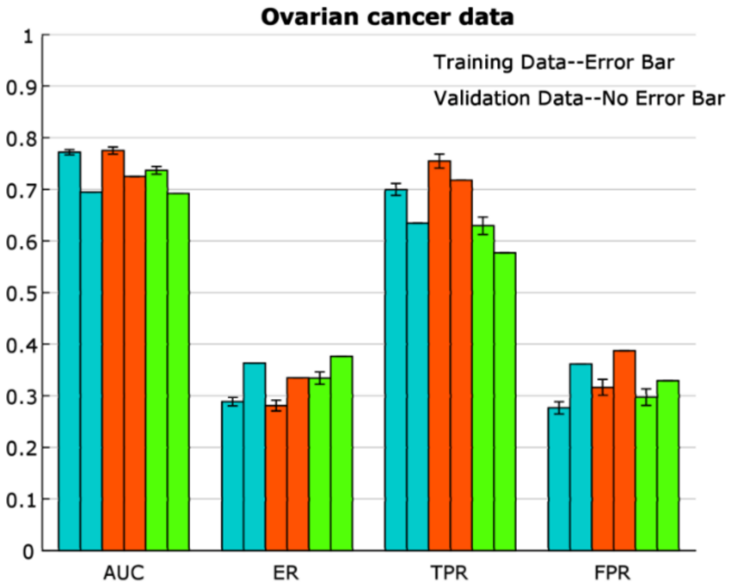


Figure 3.13 Bar plot showing performance metrics for candidate biomarkers obtained by ER (blue), TPR (orange), and FPR-estimator (green) strategy on ovarian cancer dataset. Height of first bar of each color indicates average performance during cross-validation (with error bars indicating standard deviation). The second bar indicates performance in the validation dataset.

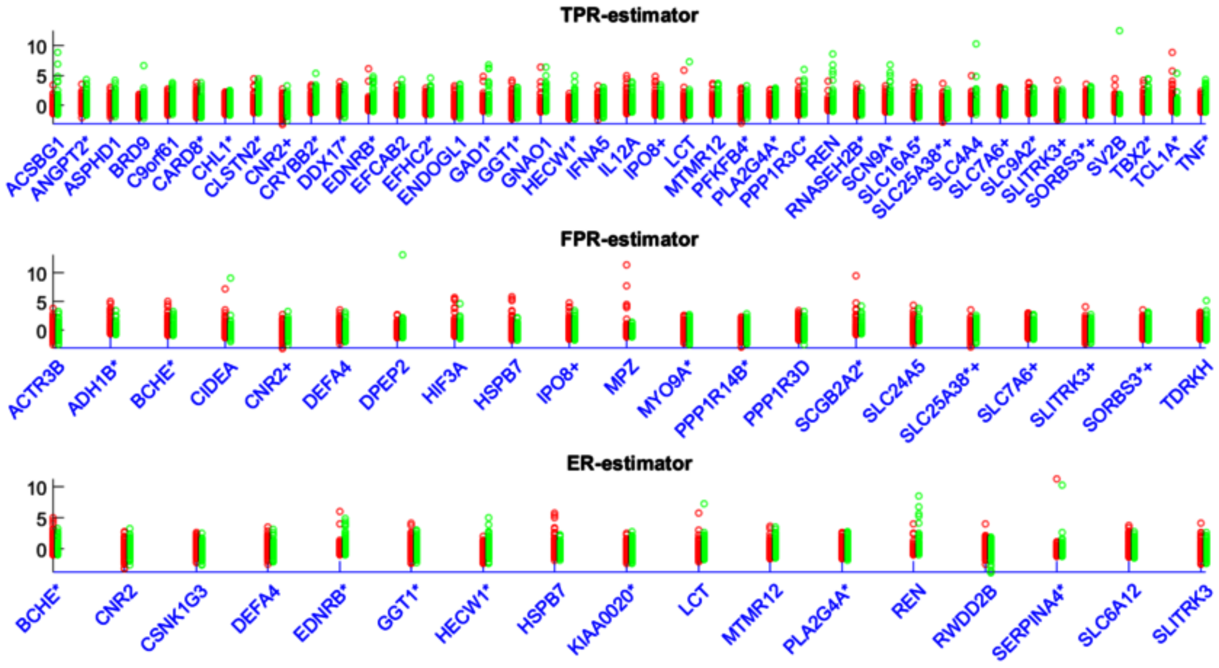


Figure 3.14 Distribution of gene expression for candidate biomarkers from the ovarian cancer dataset. Features obtained by TPR-estimator (top), FPR-estimator (middle), and ER-estimator (bottom) are shown for the positive (red) and negative class (green). Features with (*) indicate genes known to be cancerous and those with (+) indicate genes common to both TPR and FPR-estimator.

Venn Diagrams show the large fraction of candidate biomarker genes that are unique to a single estimator strategy, in addition to large fraction of known cancer genes (**Figure 3.15**). Finally, gene set enrichment analysis showed highly orthogonal pathways associated with candidate genes from each estimator strategy (**Figure 3.16**). This method may be further used to improve biomarker identification efforts in PTSD datasets, and will allow for specific tuning of True Positive or False Positive Rates.

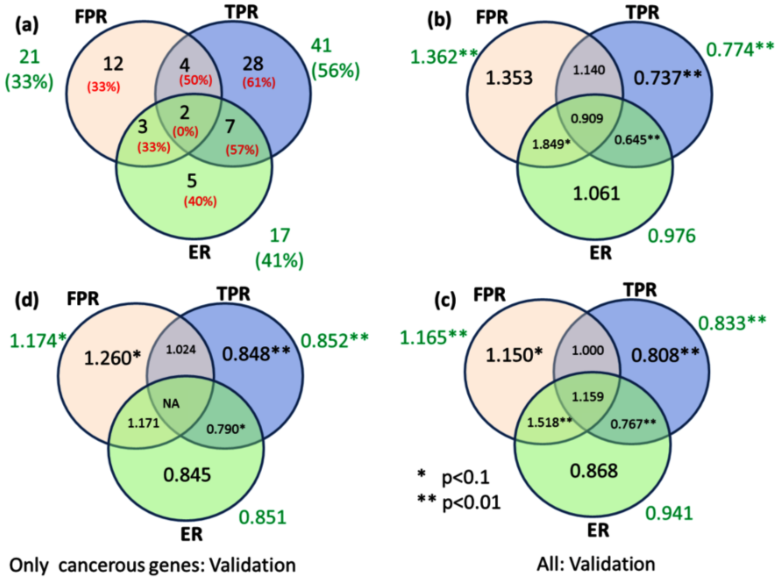


Figure 3.15 Overlap of candidate biomarkers identified by ER, TPR, and FPR-estimator strategy from ovarian cancer dataset analysis. (a) Venn diagrams indicate overlaps of identified biomarkers by each method (and fraction of genes known to be cancerous). More than half of biomarker genes identified by the TPR and FPR-estimator are unique genes, not identified by other methods. (b) Venn diagram shows the mean variance ratio (variance in positive class/variance in negative class) for each gene set, using only the cancerous genes from (a). (c) Mean variance ratio for each gene set in the validation data. Between-class variance directionality is preserved in the TPR and FPR-estimator groups, indicating stability in the identified signals. (d) Mean of variance ratio for each gene set in the validation data, using only fraction of genes known to be cancerous. Green values outside of Venn Diagram indicate the total quantities for each method.

DEVG analysis for sensitivity and specificity optimization

We have incorporated a novel feature selection algorithm in order to improve either the sensitivity or specificity of a classifier. The feature selection strategy identifies features with differing standard deviations, indicating one group shows tighter regulation while the other shows looser regulation. The looser regulation group may correspond to two (or more) tightly regulated subtypes with different mean, or a single homogeneous group with a larger variance, indicating less biological control. The notion of disease group differences in variance has been previously proposed as Differential Expression Variance Genes (DEVGs) [65]. An illustration of DEVG scenarios is show in Figure 3.17.

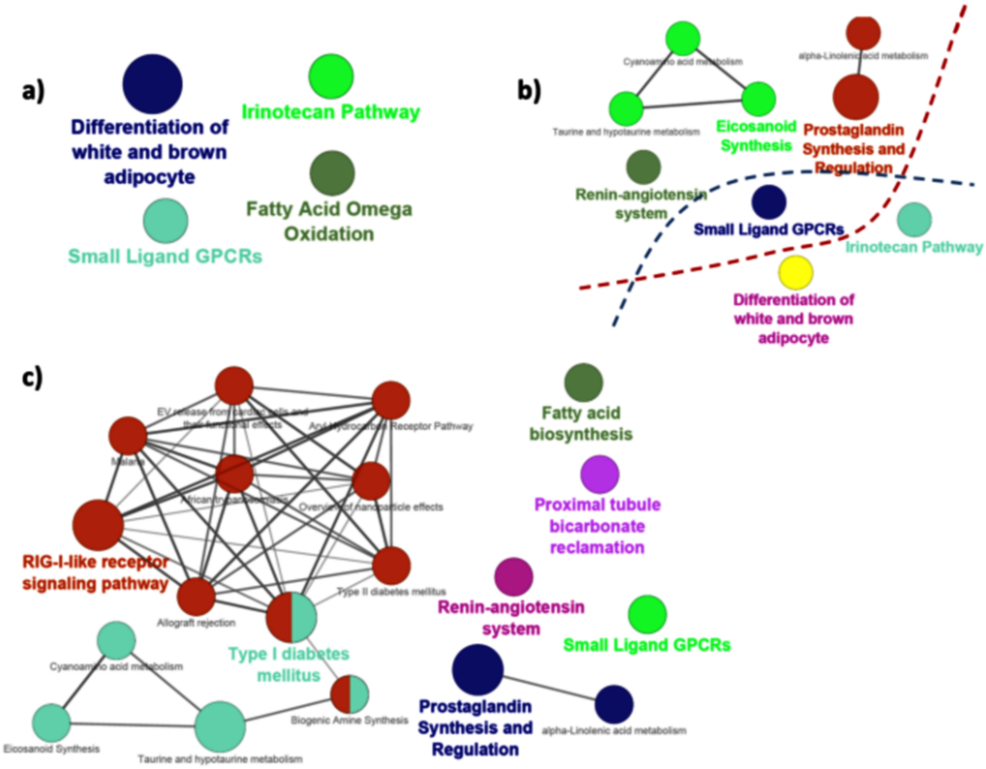

a)
Differentiation of white and brown adipocyte
Irinotecan Pathway
Fatty Acid Omega Oxidation
Small Ligand GPCRs
b)
Cyanoamino acid metabolism
alpha-Linolenic acid metabolism
Taurine and hypotaurine metabolism
Eicosanoid Synthesis
Prostaglandin Synthesis and Regulation
Renin-angiotensin system
Small Ligand GPCRs
Irinotecan Pathway
Differentiation of white and brown adipocyte
c)
EV release from cardiac cells and their functional effects
Aryl Hydrocarbon Receptor Pathway
Malaria
African trypanosomiasis
Overview of nanoparticle effects
RIG-I-like receptor signaling pathway
Type II diabetes mellitus
Allograft rejection
Type I diabetes mellitus
Cyanoamino acid metabolism
Biogenic Amine Synthesis
Eicosanoid Synthesis
Taurine and hypotaurine metabolism
Fatty acid biosynthesis
Proximal tubule bicarbonate reclamation
Renin-angiotensin system
Small Ligand GPCRs
Prostaglandin Synthesis and Regulation
alpha-Linolenic acid metabolism

Figure 3.16 Results of biomarker candidate gene enrichment analysis for ER, TPR and FPR-estimator strategy. Unique enriched pathways from FPR and TPR-estimator candidate genes are shown in (a), and (c), respectively. Enriched pathways from ER-estimator genes are shown in (b), along with pathways also identified from TPR and FPR-estimator analysis. Pathways above red dashed line in (b) were also identified by the TPR-estimator analysis, and pathways below the blue dashed line were also identified by FPR-estimator analysis.

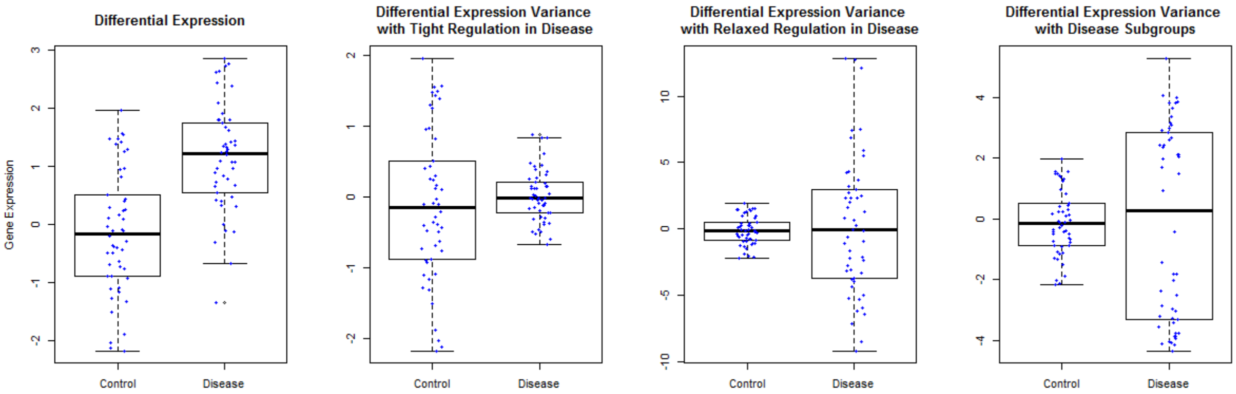


Figure 3.17 Overview of four disease gene expression distributions. (a) A traditional differentially expressed gene shows a difference in means between disease and control groups. (b)-(d) Differential expression variance identifies genes showing differential variance in expression distributions between disease and control, including cases of tight or relaxed biological control (b,c), or cases of disease subgroups which mask differences in subgroups means (d).

In our proposed methodology, we have incorporated this idea, along with traditional Differentially Expressed Genes (DEGs) to identify a robust set of features for specificity and sensitivity-specific classification. We used the following equations to rank all features for feature selection:

$$\frac{sd_{control}}{sd_{PTSD}} e^{-p} \tag{5}$$

$$\frac{sd_{PTSD}}{sd_{control}} e^{-p} \tag{6}$$

Where p is the differential expression p-value, and sd is the standard deviation of all PTSD or control samples for the gene of interest. Eqn. 5 is used to maximize the True Positive Rate (TPR), while Eqn. 6 is used to maximize the True Negative Rate (TNR).

We used the proposed feature selection algorithm along with the Nearest Shrunken Centroids (NSC) classifier to improve sensitivity and specificity. In addition to the proposed feature selection strategy, we tuned the shrinkage parameter of the NSC classifier to both minimize Error Rate and maximize the sensitivity (TPR) or specificity (TNR) during nested cross-validation. We compared the performance of the traditional classification strategy (using t-test p-values for feature selection and minimizing the error rate during centroid shrinking), with the proposed methodology using 100 rounds of 5-fold cross-validation. We applied this methodology to the metabolomics, proteomic, miRNA, endocrine, and CLIA Lab PTSD data using the 83-83 cohort for cross-validation. The proteomic dataset showed poor performance using a NSC classifier (average AUC<0.5) for traditional and modified algorithms, and has been excluded. However, the results for the four remaining datasets are shown in Figure 3.18.

In the metabolite, endocrine, and miRNA datasets, the algorithm shows minor improvements in TPR and TNR using the TPR-maximization and TNR-maximization, respectively. However, in the CLIA Lab data, our proposed algorithm does not work as expected. We do not see improvements in the TPR for the TPR-maximization algorithm, due to a loss in overall accuracy (shown by the lower AUC). We expect additional improvements can be seen with the incorporation of a TPR or TNR-specific decision boundary, or a more sophisticated classifier.

4. Biomarker identification, classification, and multi-omic analysis of PTSD datasets

4.1 Multi-omic COMBINER method for obtaining a multi-modal biomarker panel

We have extended the COMBINER platform to obtain a heterogeneous panel of candidate biomarkers. In the current strategy, we first use COMBINER to obtain single omic panels for different datasets [49]. Then, using the new proposed method (**Figure 4.1**), we integrated the single omics panels to obtain a multi-omic panel. At the integration step we select features from different modalities in order to maximize the relevance score (computed using mutual information theory), AUC, and accuracy of features on the training data. This is done by employing the greedy approach and hence does not guarantee optimal solution. Using the proposed integration method we obtained results for methylation and metabolites, miRNA and metabolites, proteins and metabolites, and proteins, metabolites and miRNA. The multi-metric performance results for these multi-omic combinations are shown in Table 4.1.

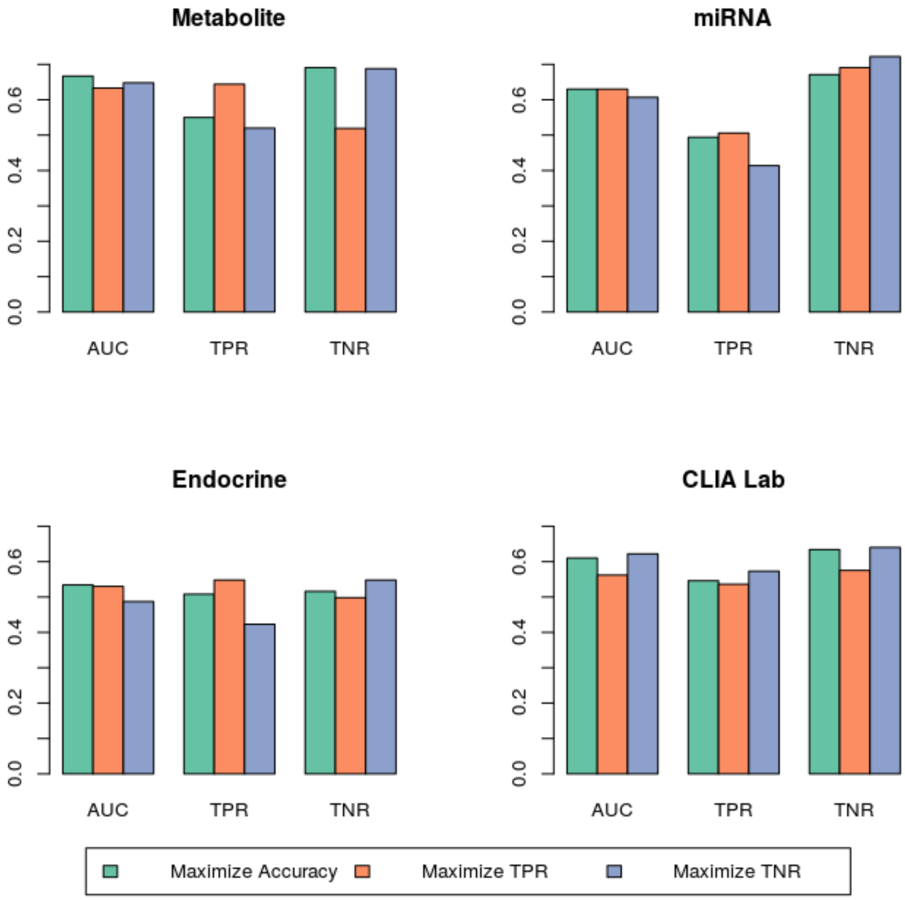


Figure 3.18 Barplot of sensitivity and specificity-optimized NSC algorithms. Height of bars indicates average cross-validated performance.

S = Set of Rank features using "Relevance" score (using Mutual Information to identify the relevance with Class labels)

STEP 1: k-fold partition of subject-matched multi-modal "discovery" set

STEP 2: Construct a subset "T" with top "t" features from ***S***
Perform cross-validation to obtain predictive AUC and ER
Find "**t**" with maximum ratio of AUC and ER

Repeat STEPS 1 and 2 500 times

STEP 3 : Find **t*** with the most robust performance i.e. consistent high ratio of AUC and ER with small variations

Figure 4.1 The proposed extension for COMBINER. First, the best feature sets for each uni-modal data type are obtained from COMBINER. These candidate uni-modal panels are concatenated and the proposed feature selection strategy is used to obtain a multi-omic candidate biomarker panel.

Table 4.1 Validation performance results obtained by COMBINER for different multi-omic datasets. We analyzed four different modalities each having at most 51 subject with PTSD and 51 subject without PTSD. Validation was performed on the 31/31 dataset.

	Methylation + Metabolite	**miRNA + Metabolite**	**Protein + Metabolite**	**Protein + miRNA + Metabolite**
AUC	0.6524	0.7519	0.6426	0.6800
ER	0.3548	0.2807	0.3929	0.3214
TPR	0.6774	0.6452	0.5806	0.6129
FPR	0.3871	0.1923	0.36	0.2400
No. of biomarkers	3	2	12	6

4.2 Semi-supervised graph-based integration

We have developed a novel multi-omics approach based on graph theory trying to integrate different data types such as DNA methylation and gene expression along with latent biological knowledge in terms of biological pathways.

Semi-Supervised Learning (SSL) methods stand between unsupervised methods, where training samples are entirely unlabeled, and supervised methods, where all training samples are labeled. SSL algorithms make use of unlabeled data along with labeled samples to enrich the training set and construct a more efficient and reliable classifier, especially when a large amount of unlabeled samples is available. The performance of such classifiers is measured on the unlabeled samples only. The key to SSL approaches is the consistency assumption which states: (1) points on the same structure (i.e., manifolds) are likely to have the same label, and (2) nearby points are also likely to have the same label. SSL methods have proven to be quite productive in dealing with complex datasets such as biological data where data structures are intertwined [66].

In this approach, each node represents a sample and the edges can be established between nodes using the *K* Nearest Neighbors (KNN) method. In fact, edges between nodes convey the mutual relationship between the samples. The more the weight of the edge, the more likely the nodes it connects to have the same label.

KNNs of each sample can be computed by ordinary Euclidean distance, and the weight of the edges obtained using the Gaussian kernel. Suppose $X = \{x_1, x_2, \ldots, x_l, x_{l+1}, \ldots, x_n\} \subset \Re^m$ to be the entire set of n samples comprising l labeled samples and $n-l$ unlabeled samples. Let $L = \{1, \ldots, c\}$ denote the class labels. Let $F = \left[F_1^t, \ldots, F_n^t\right]$ be an $n \times c$ matrix corresponding to classification of the set X, where sample x_i belongs to class j if $y_i = \arg\max_{j \in L} F_{ij}$. Here, F is a vector function which assigns a vector F_i to each sample x_i. The matrix F is obtained from Eqn. 7:

$$F = (I - \alpha S)^{-1} Y \tag{7}$$

where I is an $n \times n$ identity matrix and α denotes the tradeoff parameter between the two conditions of smoothness and loss. Also, $Y = [y_1, \ldots, y_l, 0, \ldots, 0]$ denotes the labels where samples are labeled by 1 and -1 and unlabeled samples are represented by 0. Here, $S = D - W$ is the graph Laplacian matrix, where W is the symmetric weight matrix calculated in Eqn. 8 and D is given by Eqn. 9.

$$w_{ij} = \begin{cases} \exp\left(\dfrac{\left(x_i - x_j\right)^t \left(x_i - x_j\right)}{\sigma^2} \right), & i \neq j \\ 0, & Otherwise \end{cases} \tag{8}$$

$$D = diag(d_i) \tag{9}$$

where $d_i = \sum_j w_{ij}$.

Graph integration

One of the goals of this work has been to integrate the computed graphs as the result of applying SSL on each genomic level. The purpose of graph integration is to leverage hidden knowledge in the gene expression and DNA methylation data, along with biological knowledge such as pathway information, to obtain the best classification performance. The integration process can be carried out by finding the optimal combination of each dataset represented by a graph. Suppose that there are K graphs. The weights of the combined graphs can be obtained using the following optimization model:

$$\begin{aligned} &\min_{\alpha} Y^t \left(I + \sum_{k=1}^{K} \alpha_k S_k \right)^{-1} Y \\ &s.t. \quad \sum_{k=1}^{K} \alpha_k \leq \mu \end{aligned} \tag{10}$$

where S_k and α_k represent the graph Laplacian matrix and the optimum weight coefficient of the graph k, respectively.

The final solution of the abovementioned model can be calculated by Eqn. 11:

$$F = (I - \sum_{k=1}^{K} \alpha_k S_k)^{-1} Y \tag{11}$$

The approach

The main contribution of this work has been to provide a systematic approach to incorporate biological knowledge in the form of biological pathways into a graph-based SSL algorithm, to gain a better phenotype classification performance. Each genomic level such as gene expression or DNA methylation graph has a complementary graph containing its corresponding pathway information. Figure 4.2 illustrates the overall pipeline of the approach. In this figure, each node represents a sample, where the samples are the same for all genomic levels being considered. Graphs of each level are constructed by the SSL algorithm. It should be mentioned that in the context of SSL, all samples including labeled and unlabeled are taken into account during the process of learning. In Figure 4.2 the two-class problem is addressed where node classes are represented by ‘1’ and ‘0’ and unlabeled samples are represented by ‘?’.

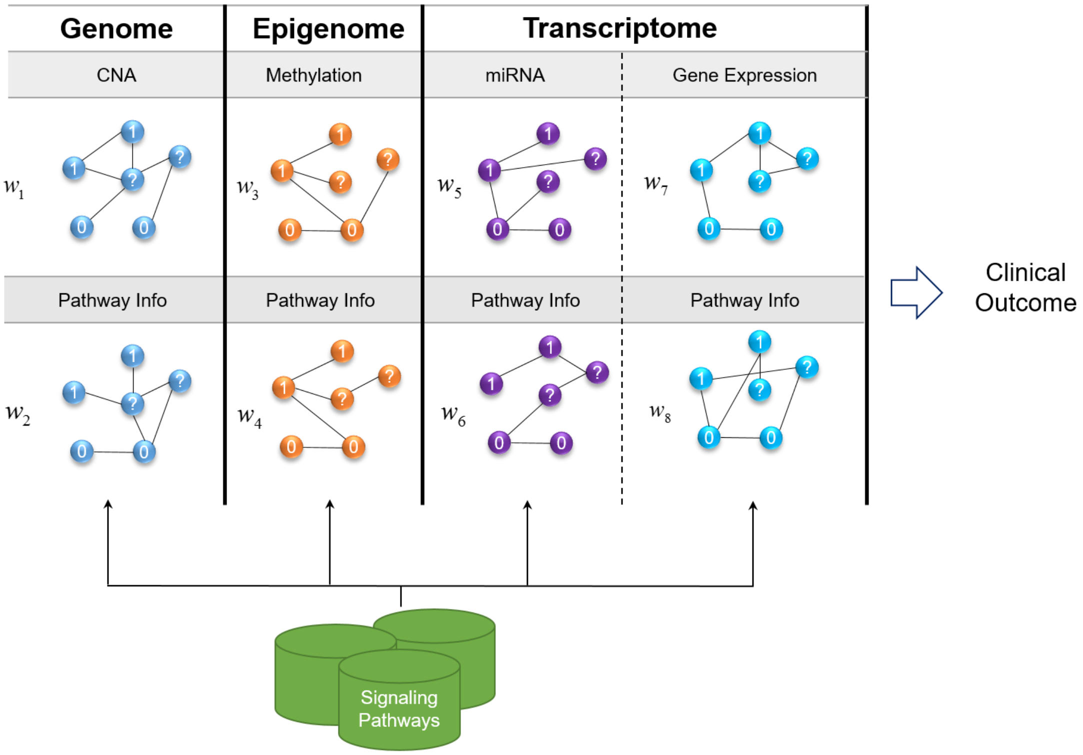


Figure 4.2 A graphical representation of the graph integration method.

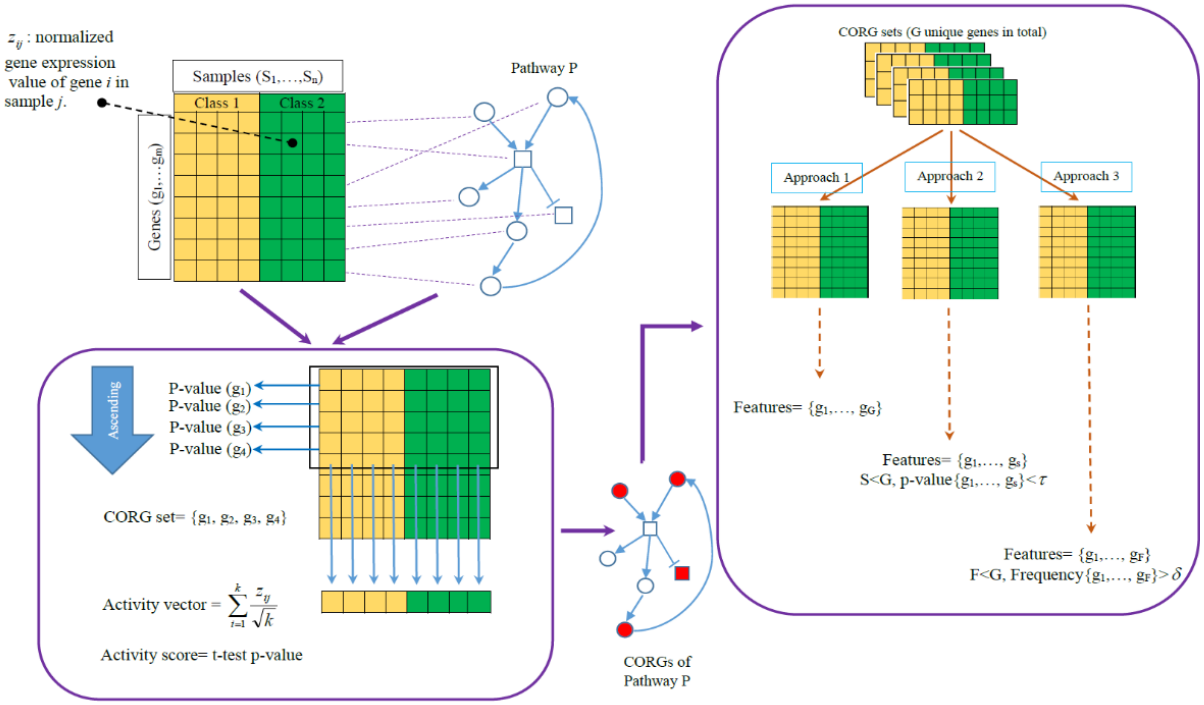

z_{ij} : normalized gene expression value of gene i in sample j.
Samples ($S_1,\ldots,S_n$)
Class 1
Class 2
Genes ($g_1,\ldots,g_m$)
Pathway P
Ascending
P-value (g_1)
P-value (g_2)
P-value (g_3)
P-value (g_4)
CORG set= {g_1, g_2, g_3, g_4}
Activity vector = $\sum_{i=1}^{k} \frac{z_{ij}}{\sqrt{k}}$
Activity score= t-test p-value
CORGs of Pathway P
CORG sets (G unique genes in total)
Approach 1
Approach 2
Approach 3
Features= {$g_1,\ldots, g_G$}
Features= {$g_1,\ldots, g_S$}
S<G, p-value{$g_1,\ldots, g_S$}<τ
Features= {$g_1,\ldots, g_F$}
F<G, Frequency{$g_1,\ldots, g_F$}>δ

Figure 4.3 Gene extraction process from biological pathways.

Although construction of the graphs with respect to each dataset is performed by the existing SSL algorithm discussed in the Section 1.2.1, in order to construct the graphs considering biological pathways, three new approaches have been developed based on the set of COndition Responsive Genes (CORGs) [67]. After extracting the entire signaling genes (CORGs) of each pathway, three approaches were considered to shape the final set of genes for constructing the pathway graphs. In Approach 1, all the genes in all the CORGs were listed and used as the final set of features. Note that it is possible that some genes are repeated in various CORGs. In such cases, just one of them is adopted. In Approach 2, all the unique genes in the CORGs are ordered in an ascending manner based on their p-values. Then, genes with p-values larger than a threshold are filtered out. The threshold that we have set in this work for filtering genes was 0.001. Finally, in Approach 3, we make use of the number of times that each gene has been repeated

in the CORGs. The more a gene is repeated, the stronger it is as a biomarker. Finally, a threshold is applied and low frequency genes are filtered. In this paper, we set the threshold to be 0.01x(# pathways). The overall pipeline of these approaches is depicted in Figure 4.3. The process of biomarker discovery is done through a greedy intelligent search based on Simulated Annealing meta-heuristic search algorithm. This process is demonstrated in Figure 4.4.

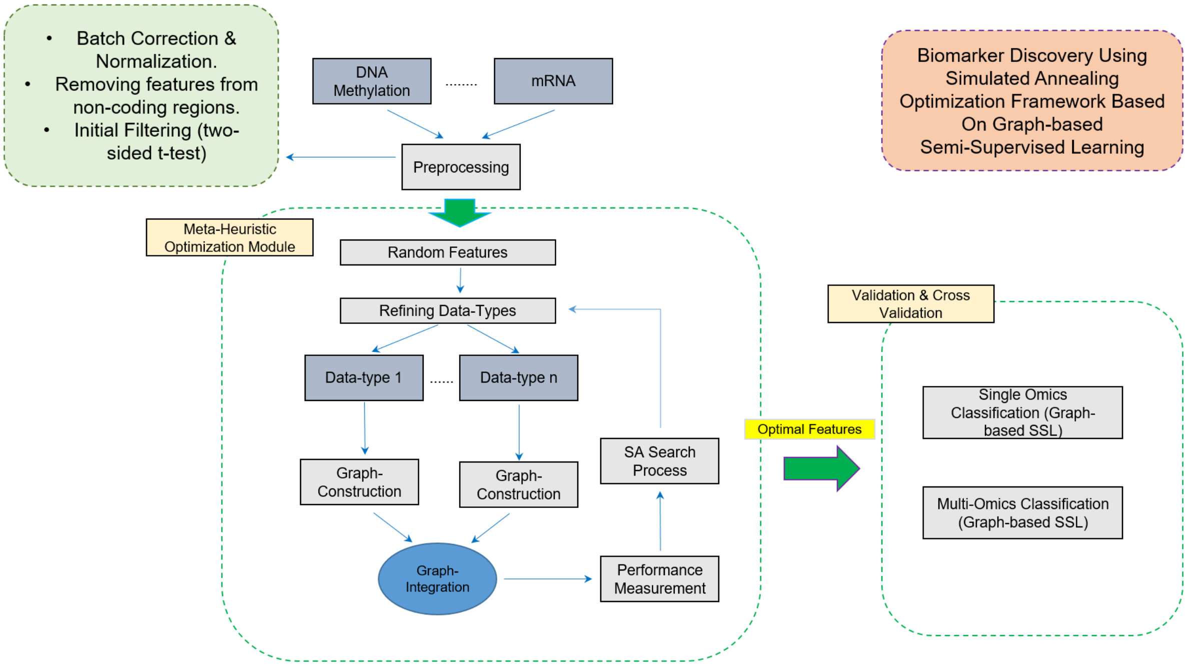

• Batch Correction & Normalization.
• Removing features from non-coding regions.
• Initial Filtering (two-sided t-test)
DNA Methylation
....
mRNA
Preprocessing
Biomarker Discovery Using Simulated Annealing Optimization Framework Based On Graph-based Semi-Supervised Learning
Meta-Heuristic Optimization Module
Random Features
Refining Data-Types
Data-type 1
......
Data-type n
Graph-Construction
Graph-Construction
Graph-Integration
Performance Measurement
SA Search Process
Optimal Features
Validation & Cross Validation
Single Omics Classification (Graph-based SSL)
Multi-Omics Classification (Graph-based SSL)

Figure 4.4 Biomarker discovery pipeline in the proposed graph-based SSL.

In what follows, we have represented single and multi-omics experimental results along with the obtained sets of biomarkers (Figures 4.5-4.6, Tables 4.2-4.5).

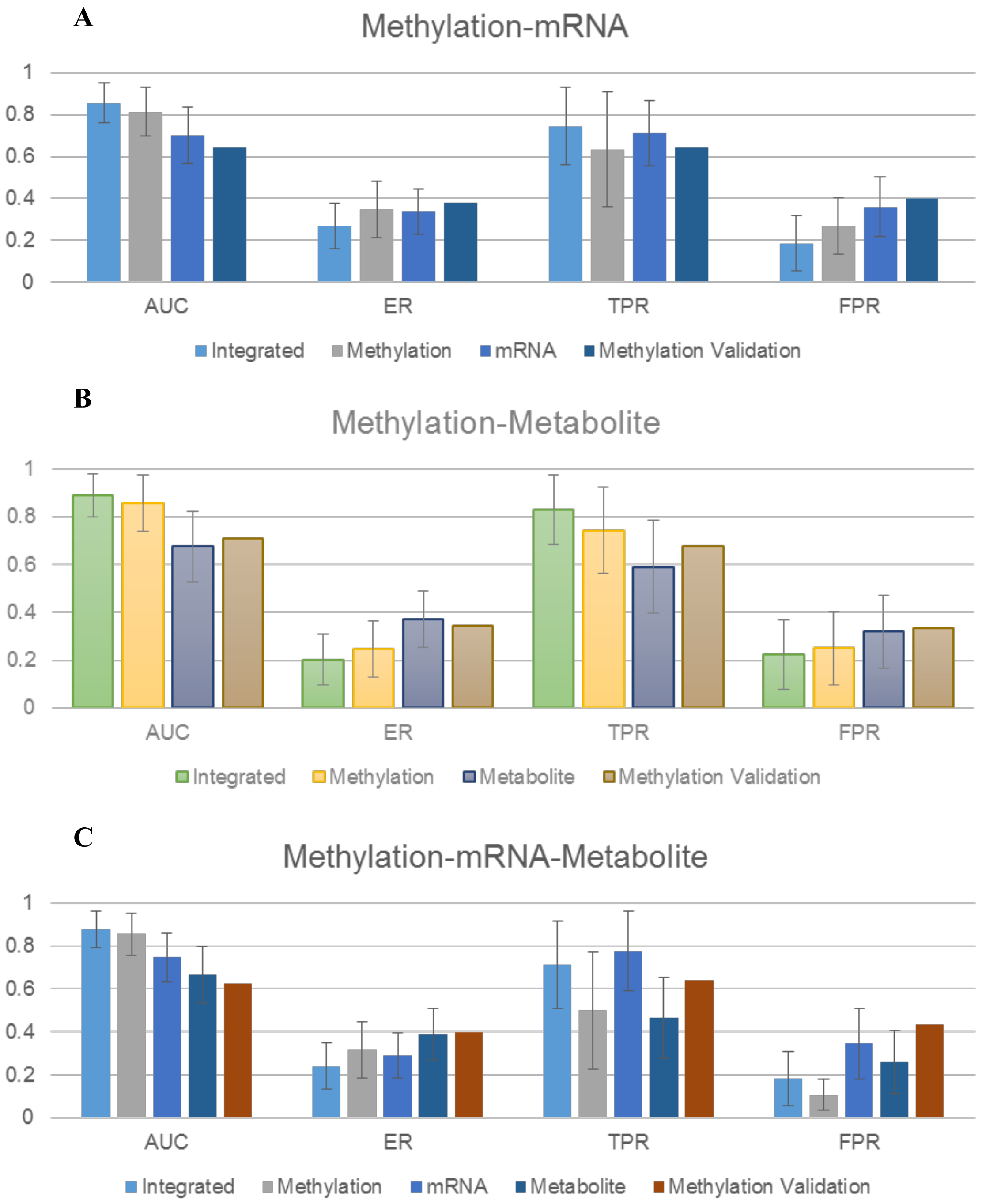


Figure 4.5 Biomarker performance from graph-based integration of (A) Methylation and mRNA, (B) Methylation and Metabolite, and (C) Methylation, mRNA, and Metabolite.

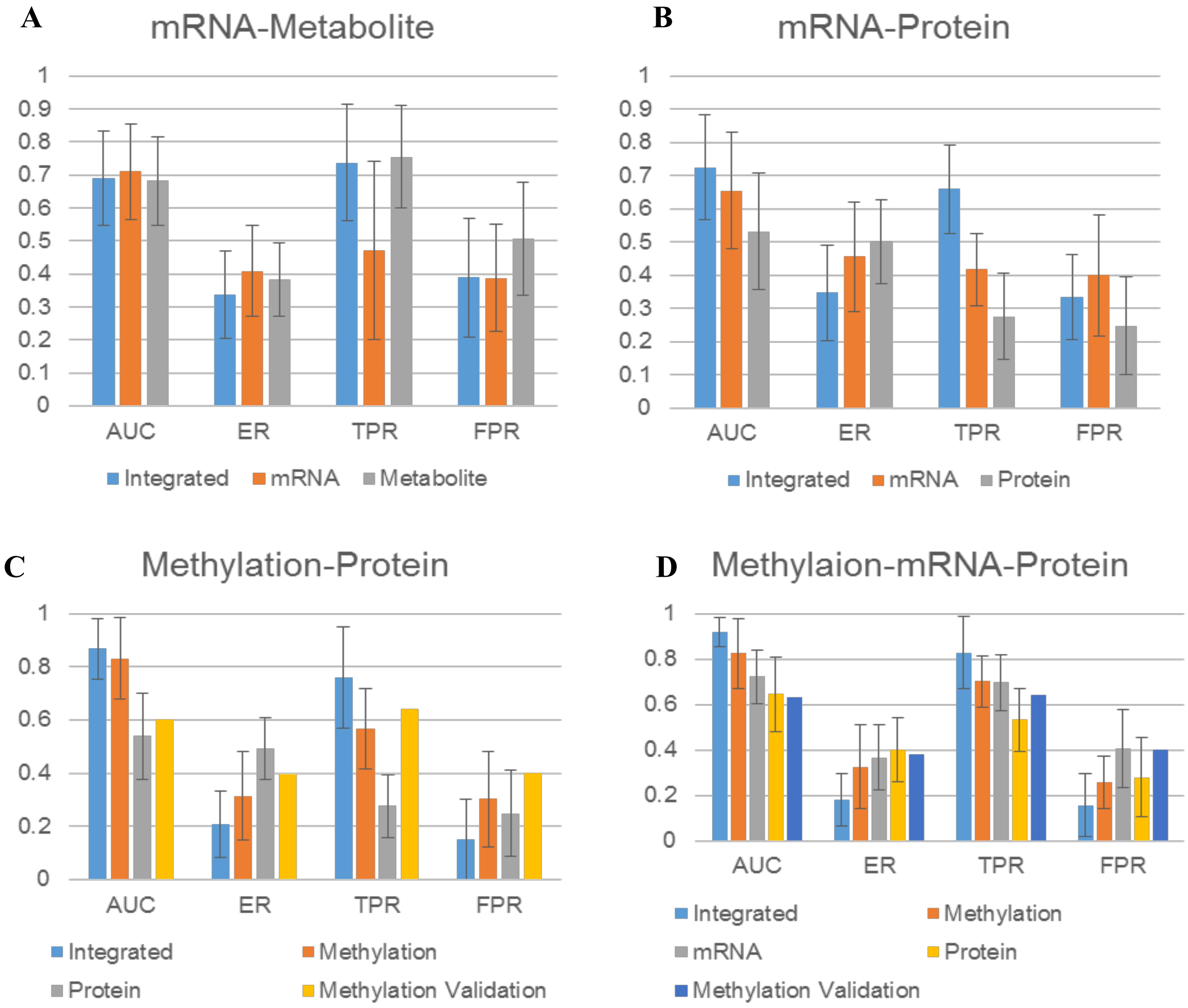


Figure 4.6 Biomarker performance from graph-based integration of (A) mRNA and Metabolite, (B) mRNA and Protein, (C) Methylation and Protein, and (D) Methylation, mRNA and Protein.

Table 4.2 Biomarker list of the DNA methylation data (Illumina 83-83)

CpG Site	Gene	p-value
cg14080518	SMURF1	8.36E-06
cg23131950	AP2S1	9.41E-06
cg00022594	AKAP8L	1E-05
cg22661330	LANCL2	1.57E-05
cg05452391	C5orf56	2.71E-05
cg02779164	SETBP1	0.00905
cg14202338	GIPC3	0.007265
cg04353053	C1orf127	0.007703
cg01302119	WDR60	0.008452
cg19889580	KCTD16	0.002264
cg10525567	NCKIPSD	0.009311
cg23968383	ZNF572	0.00354
cg16055159	CEP97	0.000887
cg07502936	ZNF135	0.006944
cg12154261	TDRD9	0.001504
cg02992296	INTS1	0.006454
cg24642820	NUP210	0.007784
cg12628062	PSMC3IP	0.005791
cg14474728	RPH3AL	0.003815
cg24037389	SCARA5	0.009535
cg09267483	BTBD17	0.007057
cg20436533	CDK15	0.005272
cg24508208	CALD1	0.000456
cg26489108	DMRT3	0.003994
cg10202544	TC2N	0.003754
cg14596589	SLC12A7	0.001306
cg14492241	SNX8	0.0034
cg17562528	DSCAM	0.000432
cg12476052	SV2B	0.009245
cg13861527	BRE	0.009658
cg02695697	MMP16	0.00488
cg12385425	TMEM17	0.007685
cg10332704	THEM5	0.009413
cg19520337	CPNE8	0.009181

Table 4.3 Biomarker list of the metabolite data (83-83).

Biomarker	**p-value**
adrenate224n6	0.534074
phenylacetylglutamine	0.196939
stachydrine	0.172708
ifn_g	0.705392
m2hydroxypalmitate	0.006966
m5alphapregnan3beta20alphadioldisulfate	0.139472
palmitate160	0.162172
alphahydroxyisovalerate	0.171466
methyl4hydroxybenzoate	0.512394
urea	0.407291
dodecanedioate	0.313245
dihomolinoleate202n6	0.003027
m1stearoylglycerol	0.440302
m4androsten3beta17betadioldisulfate1	0.652193
sarcosineNMethylglycine	0.073874
Nacetylglycine	0.302841
phenyllactatePLA	0.044491
creatinine_metabolon	0.711369
m3methyl2oxovalerate	0.288728
gammaglutamylalanine	0.157785
c45_l_3n56tp16tp_16p56n	0.003636
gsh_gssg	0.115121
m2methoxyacetaminophensulfate	0.579075
m13dipalmitoylglycerol	0.006984
erythrosphingosine1phosphate	0.052747
tnf_a	0.182351
salicylate	0.368304

Table 4.4 Biomarker list of the mRNA data (83-83).

Biomarker	Gene	p-value
A_23_P253395	UNC13B	0.0000106
A_23_P46936	EGR2	0.0000156
A_23_P214080	EGR1	0.0000236
A_23_P62901	BTG2	0.0000375
A_24_P144773	RNF145	0.005245
A_33_P3308045	EIF4E2	0.002587
A_23_P302060	IFNE	0.007967
A_33_P3381305	NA	0.00745
A_33_P3669411	NA	0.000302
A_33_P3421664	TDRD5	0.002186
A_23_P500861	SYNE1	0.008167
A_33_P3777165	FLJ31715	0.004973
A_33_P3397613	NA	0.004402
A_33_P3441060	C6orf35	0.003613
A_33_P3318292	SFPQ	0.007425
A_23_P79518	IL1B	0.000159
A_33_P3417620	ZRSR2	0.001234
A_23_P123539	PPP2R2A	0.007501
A_24_P658584	SASH1	0.0035
A_24_P16124	IFITM4P	0.001114
A_24_P81900	SLC2A3	0.006441
A_33_P3364864	NAMPT	0.003411
A_33_P3238007	LARP1B	0.002119
A_33_P3276386	NA	0.001686
A_23_P323166	SRRM2	0.008381
A_33_P3402489	OAS3	0.007179

Table 4.5 Biomarker list of the protein data (83-83).

Biomarker	p-value
PON1.IQN	0.004135
PPBP.ICL	0.009251
IGFALS.DFA	0.009404
PON1.IFF	0.012719
APOA2.SPE	0.423942
PPBP.GTH	0.044186
A1BG.SGL	0.804879
HABP2.LIA	0.506517
C8B.IPG	0.38084

4.3 Greedy Multi-view Learning for Multi-Omics Data Analysis

In this project, we have developed a multi-omics classification and biomarker discovery approach based on the concept of multi-view learning. The overall pipeline of this method is depicted in Figure 4.7.

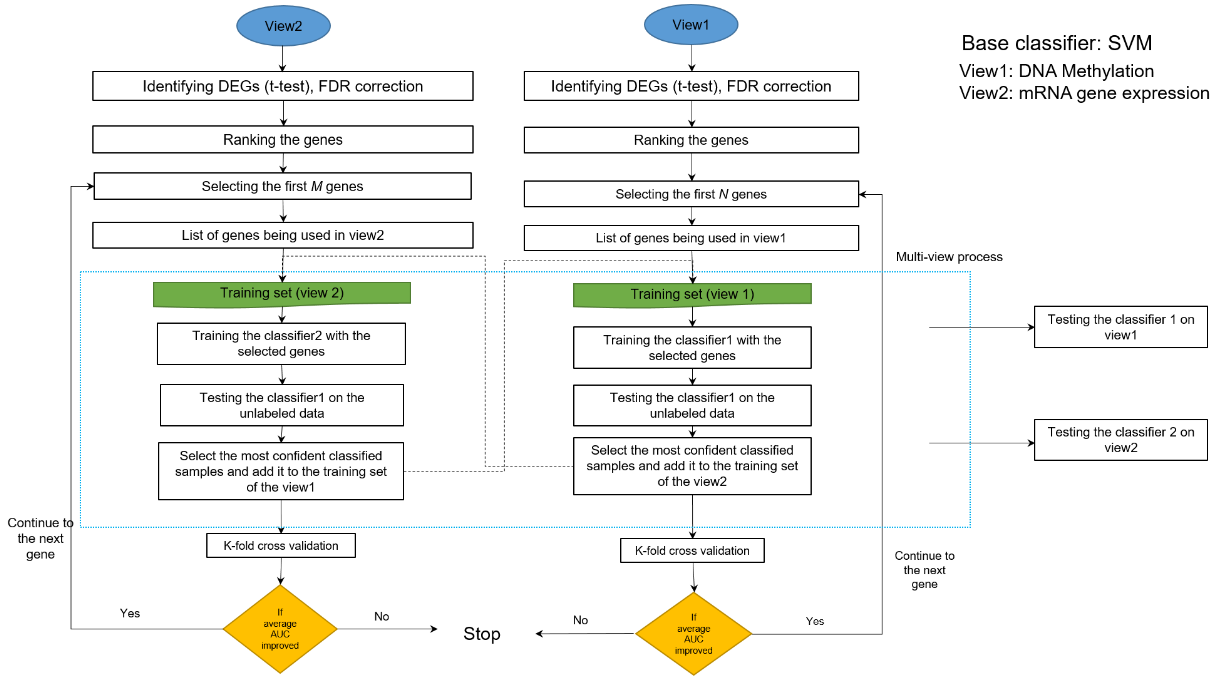

View2
View1
Base classifier: SVM
View1: DNA Methylation
View2: mRNA gene expression
Identifying DEGs (t-test), FDR correction
Identifying DEGs (t-test), FDR correction
Ranking the genes
Ranking the genes
Selecting the first *M* genes
Selecting the first *N* genes
List of genes being used in view2
List of genes being used in view1
Multi-view process
Training set (view 2)
Training set (view 1)
Training the classifier2 with the selected genes
Training the classifier1 with the selected genes
Testing the classifier 1 on view1
Testing the classifier1 on the unlabeled data
Testing the classifier1 on the unlabeled data
Testing the classifier 2 on view2
Select the most confident classified samples and add it to the training set of the view1
Select the most confident classified samples and add it to the training set of the view2
Continue to the next gene
K-fold cross validation
K-fold cross validation
Continue to the next gene
Yes
If average AUC improved
No
Stop
No
If average AUC improved
Yes

Figure 4.7. Multi-view learning and biomarker discovery.

The numerical experiments were performed on the Agilent gene expression and methylation datasets, using the 52-52 for training and 31-31 for validation. The results of the DNA methylation validation are represented in Table 4.6 and a list of the identified CpG sites are reported in Table 4.7.

Table 4.6 Classification performance for the greedy multi-view learning approach on Agilent DNA methylation data.

	Greedy multi-view method	
Metric	P	A
AUC	0.622	0.523
ER	0.41	0.491
MCC	0.139	0.026
MSPE	0.225	0.278
Youden	0.341	0.175
PPV	0.5881	0.511
NPV	0.592	0.530
TPR	0.572	0.512
FPR	0.377	0.441

Table 4.7 The biomarker CpG sites identified by the multi-view greedy search algorithm.

ID	Symbol	Name	p-value	FDR
A_17_P15535760	MFSD10	Encodes a member of the major facilitator superfamily of transporter proteins.	2.017E-08	0.0096
A_17_P16992190	KPNB1	Diseases associated with KPNB1 include Campomelic dysplasia, a severe disorder that affects the development of the skeleton and reproductive system.	2.637E-08	0.0153
A_17_P15006698	SAMD11	May play a role in photoreceptor development.	4.992E-08	0.0102
A_17_P15229454	WDR43	A Protein Coding gene.	8.152E-08	0.0226
A_17_P11194228	ZMYND8	This gene encodes a receptor for activated C-kinase (RACK) protein.	1.235E-06	0.0139
A_17_P16821815	PDE8A	The protein encoded by this gene belongs to the cyclic nucleotide phosphodiesterase (PDE) family, and PDE8 subfamily.	20965E-06	0.0614
A_17_P32097142	DGCR6	Could play a role in neural crest cells migration.	3.876E-06	0.0265
A_17_P02004261	KIF1A	a motor protein involved in the anterograde transport of synaptic-vesicle precursors along axons.	1.94E-06	0.0472
A_17_P03396998	ANXA5	Diseases associated with ANXA5 include pregnancy loss and amelanotic melanoma.	1.917E-06	0.055
A_17_P16378261	LZTS2	Negative regulator of the Wnt signaling pathway.	1.847E-06	0.0069

4.4 Boosted decision trees for multi-omic classification with missing samples

To integrate multi-omic datasets, we have developed a confidence-based boosted decision tree classifier. Feature selection, tree building, and final classifier construction are completed independently on each available data type using the 'gbm' package in R. Predictions are made for each data type in the validation dataset, including the probability of belonging to either the PTSD or control class. Finally, for subject with multiple available datasets, the most extreme probability (closest to 0 or 1) over all datasets is used to determine the disease status. Using this strategy, we compare all available datasets but use only one dataset for each subject to make the final prediction. This allows subjects to classified based on the "best" dataset for each person. This strategy may help to reduce the effect of noise on predictions and may help capture the existence of subtypes in some (or all) molecular data types. Additionally, by using a decision tree-based approach, missing values can be easily accommodated by sending missing data to a third node at each split. Example trees illustrating this are shown in Figure 4.8.

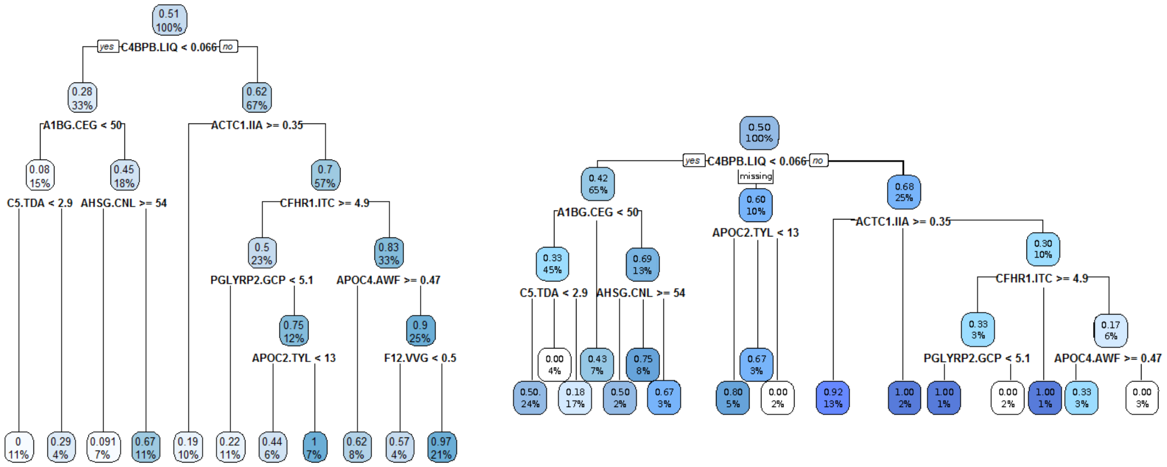


Figure 4.8 Diagram of decision tree incorporation of missing data. Left: Traditional decision tree without missing data. Each node in the tree is divided by the split point of the designated variable into two branches. Each node is labeled with the fraction of disease samples and percentage of overall data on that branch. Each variable and split point is chosen to maximize the purity of the branches below. Right: Decision tree with missing data. Instead of two branches emerging from each node, one for above and one for below the split point, three branches are shown. The third branch contains all subjects for which the chosen variable was missing.

Table 4.8 shows a summary of the available 83/83 molecular data used to evaluate this method. In particular, by incorporating the subjects with missing mRNA and miRNA data, the statistical power can be improved for classification.

Table 4.8 Summary of available 83/83 data for multi-omic classification.

Data Type	# of features used for classification	# of PTSD Subjects	# of Control Subjects
DNA Methylation	429948	81	82
mRNA	50599	76	80
Metabolite	244	82	83
Protein	96	82	80
Endocrine	35	82	83
CLIA Lab	44	81	82
miRNA	43	71	74

To compare the proposed multi-omic classification strategy with the performance of each individual data type, we performed an identical feature selection strategy, and decision tree classifier to each data type individually. The performance of each data type over 100 rounds of 5-fold cross-validation is shown along with the multi-omic performance for comparison. The average performance over all cross-validation runs is shown in Table 4.9. Integrating additional data types results in similar or better performance than the best included single data type for all considered multi-omic combinations. The best performing multi-omic combination resulted in an average AUC of 0.714 by integrating Clinical Lab data (CLIA), metabolites, and miRNAs. This integrated result was higher than the best of those three data types independently (0.696 AUC for miRNA).

Table 4.9 Summary of single and multi-omic classification performance. Classification performance metrics over 100 rounds of cross-validation are summarized by mean ± standard deviation.

	AUC	**ER**	**Sensitivity (TPR)**	**Specificity (TNR)**
Methylation	0.583±0.13	0.442±0.09	0.608±0.19	0.508±0.20
mRNA	0.494±0.16	0.507±0.12	0.500±0.18	0.486±0.18
Metabolite	0.690±0.10	0.370±0.10	0.580±0.15	0.680±0.15
Protein	0.571±0.14	0.411±0.12	0.563±0.18	0.615±0.18
miRNA	0.696±0.12	0.347±0.11	0.706±0.14	0.600±0.21
Endocrine	0.5843±0.12	0.451±0.12	0.545±0.19	0.552±0.15
CLIA Lab	0.665±0.12	0.379±0.10	0.580±0.17	0.662±0.15
CLIA + metabolite	0.696±0.13	0.331±0.11	0.613±0.16	0.732±0.16
CLIA + metabolite + miRNA	0.714±0.10	0.320±0.10	0.675±0.16	0.684±0.12
Methylation + mRNA	0.563±0.15	0.487±0.13	0.422±0.16	0.615±0.21
Protein + metabolite + endocrine	0.693±0.10	0.353±0.11	0.618±0.18	0.680±0.11

4.5. Integrating clinical and molecular data for improved classification performance

To improve biomarker identification and classification performance, we have previously integrated molecular datasests with clinical, demographic, and physiological data to create a hybrid biomarker panel for further validation. Using a decision tree classifier (which does not require normalization between molecular and clinical features), each molecular dataset was merged with potential clinical datasets (excluding CAPS and other PTSD-defining variables). These hybrid datasets were used to train and test decision tree classifiers using nested cross-validation on the male 83-83 discovery cohort. The boosted decision tree classifier was implemented using the 'gbm' package in R. A summary of the hybrid panel performances using molecular and physiological data is shown in Table 4.10.

Table 4.10 Summary of integrated physiological and molecular dataset classification performance.

Molecular Dataset	Cross-validated AUC in 52-52 discovery dataset	Test AUC in 31-31 dataset	Physiological markers identified	Molecular markers identified
Methylation	0.530 ± 0.188	0.416	• Systolic blood pressure • Diastolic blood pressure • height	• cg23771949 • cg03890840 • cg00001583 • cg00007036
Metabolite	0.572 ± 0.181	0.677	• height • weight	• Gamma-glutamylisoleucine* • Pro-hydroxyl-pro
miRNA	0.674 ± 0.181	0.630	• bmi • systolic blood pressure	• miR_33a • miR_421 • miR_146b • miR_382 • miR_374a
Protein	0.642 ± 0.190	0.489	• systolic blood pressure • Waist circumference	• GSTO1-GSA • SERPINA10-IFS • PPBP-ICL • PTGDS-AQG

5. Identification and characterization of human PTSD subtypes from DNA methylation data

We have continued to use unsupervised learning techniques to identify molecularly-defined subtypes of PTSD. To show agreement among many levels of biological data, we created a "subtype agreement matrix" to quantify the overlap of PTSD subtypes signals.

Two PTSD subtypes were identified from promoter-region DNA Methylation patterns. To reduce the noise, probe-level methylation patterns were projected onto known biology pathways using *pathifier* [68]. This reduced the dimensionality of the data from approximately 150,000 probes to 1320 compute pathway activity scores. Next, hierarchical and model-based clustering were used to identify subtypes from the data. A heatmap of pathway activity scores for all PTSD subjects is shown in Figure 5.1, including a hierarchical clustering dendrogram indicating the subtypes. Model-based clustering of the pathway activity scores identified two PTSD subtypes, containing 36 and 45 subjects.

To quantify the agreement between these subtypes identified from DNA methylation patterns and other molecular data types (mRNA, protein, metabolite, etc), we performed differential expression analysis between PTSD subgroups. Each data type was independently used to identify subtypes, and all remaining datasets were used to quantify subtype signal. The heatmap in Figure 5.2 shows the similarity and strength of PTSD subtype signals across all molecular data types. Although there is not strong agreement between all data types, the heatmap clearly shows lack of agreement if mRNA, Endocrine, or Clinical data is used to define the subtypes (indicated by close to 0% DEGs in all other data types). Additionally, clusters identified from all methylation probes show good agreement with cluster identified from probes only in the promoter region, indicating a location-independent epigenetic signal.

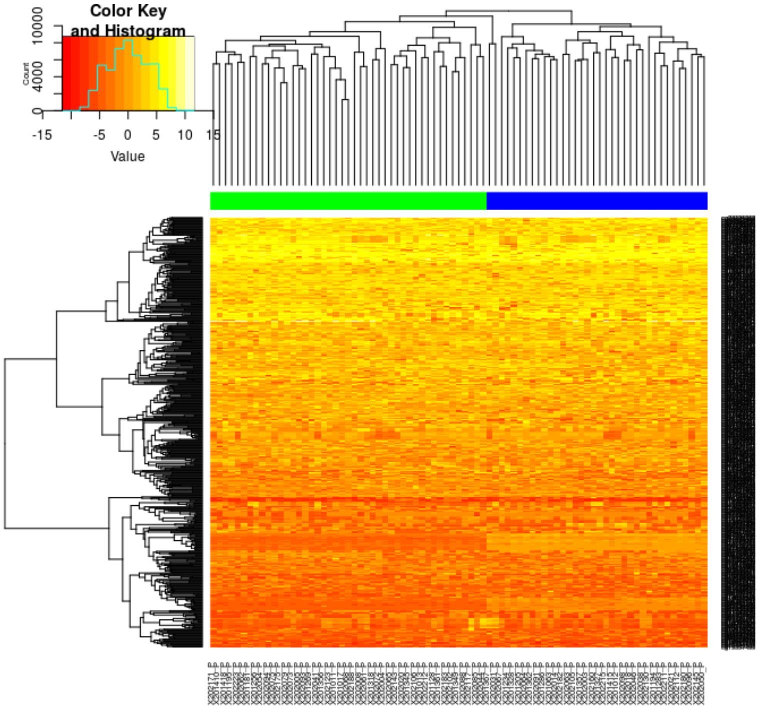


Figure 5.1 Heatmap and dendrogram of DNA Methylation patterns defining PTSD subtypes. Columns represent individual subjects, while rows indicate the methylation patterns of a specific CpG probe. The color indicates levels of methylation. The blue and green color bar along the top dive the PTSD into subjects into Subtype 1 (blue) and Subtype 2 (green).

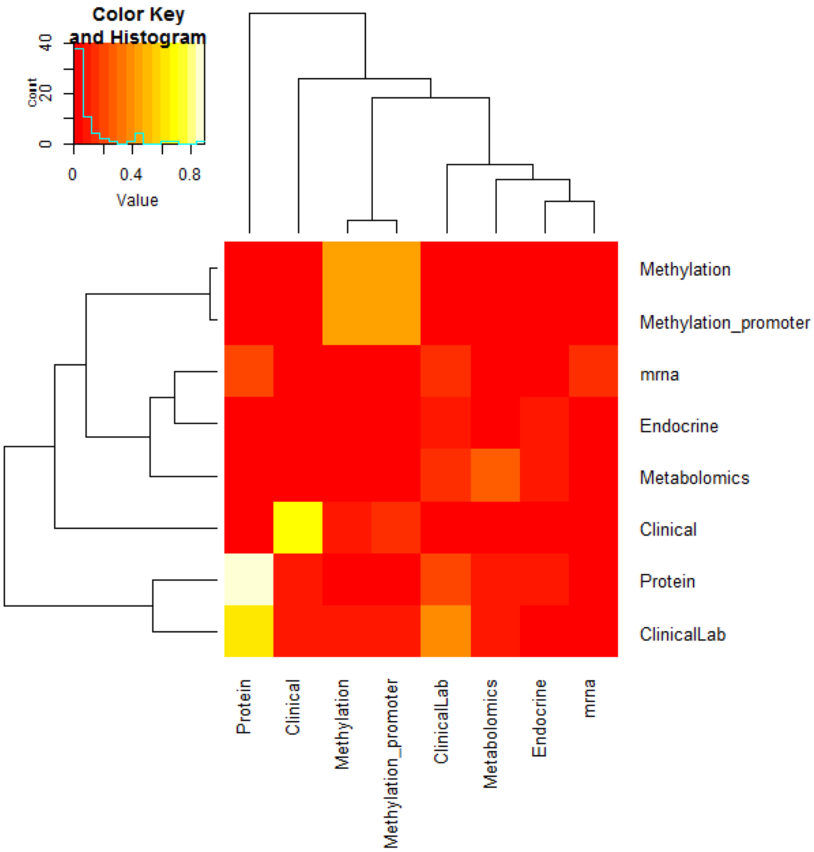


Figure 5.2 Heatmap of PTSD subtype DEG signals. Two PTSD subgroups were identified from the dataset indicated by each column label. The fraction of differentially expressed features between the identified subtypes is computed in all datasets (indicated by the row label).

Due to the long term stability of DNA methylation, and literature suggesting trauma may result in DNA methylation changes [17], we selected DNA methylation as the primary dataset for PTSD subgroup identification. In both the dataset containing all methylation probes, and the promoter region only subset, approximately 40% of CpG sites were differentially methylated between PTSD Subtype 1 and 2 (uncorrected $p<0.01$). The dataset showing the next strongest signal between these two subtypes was the Clinical dataset, indicating subtypes differences in PTSD symptom severity (CAPS, PTSD Checklist Score, etc).

6. Development and analysis of an HPA-Circadian-Metabolic model

Cortisol is a neuroendocrine hormone of Hypothalamus-Pituitary-Adrenal (HPA) axis, known to oscillate in a circadian manner. Stress modulates cortisol levels to maintain homeostasis by counteracting the stress effect by regulating the metabolic and neurological processes. Disturbances in cortisol peak and nadir levels, period, and amplitude are known to contribute to several neuropsychiatric diseases. In this study we focused on the cortisol profiles in major depressive disorder (MDD) and Post-traumatic Stress Disorder (PTSD). A systems-level perspective of cortisol dynamics is essential to analyze the underlying mechanisms that shape the cortisol profiles. We analyzed a detailed model incorporating the circadian mechanisms to characterize cortisol profiles in healthy, MDD and PTSD subjects.

6.1 Characterization of cortisol profiles and Model Development

The data for cortisol profiles in three subjects each, for healthy, MDD and PTSD phenotypes was published by Yehuda et al. [69] and was modeled and analyzed by Sriram et al. [70]. The simulated cortisol profiles for 50 sets of model parameters had shown statistically significant difference in the peak and nadir levels in the healthy, MDD and PTSD subjects. We used the mean peak and nadir values of the cortisol profiles to characterize the disease phenotypes in the simulations (**Figure 6.1**). We developed an integrated model from published literature [70-74] with submodules for the SCN clock, HPA axis and adrenal clock. The model consists of 60 ODEs and 235 parameters. The detailed model was reparametrized after integration and simulated using MATLAB. The parameters for steroidogenesis in the adrenal gland were obtained from the data reported by Son et al. [74]. Figure 6.2A shows model simulations for the circadian dynamics of the SCN clock, cortisol, adrenal clock and steroidogenesis (star protein). The model was tested for obtaining qualitative profiles and appropriate fold changes of cortisol under normal and mutant conditions to verify the cortisol dynamics as reported in experimental observations. Figure 6.2B depicts the scenarios for the effect of ablation of SCN drive, ablation of adrenal clock and ablation of both the clocks on cortisol profiles.

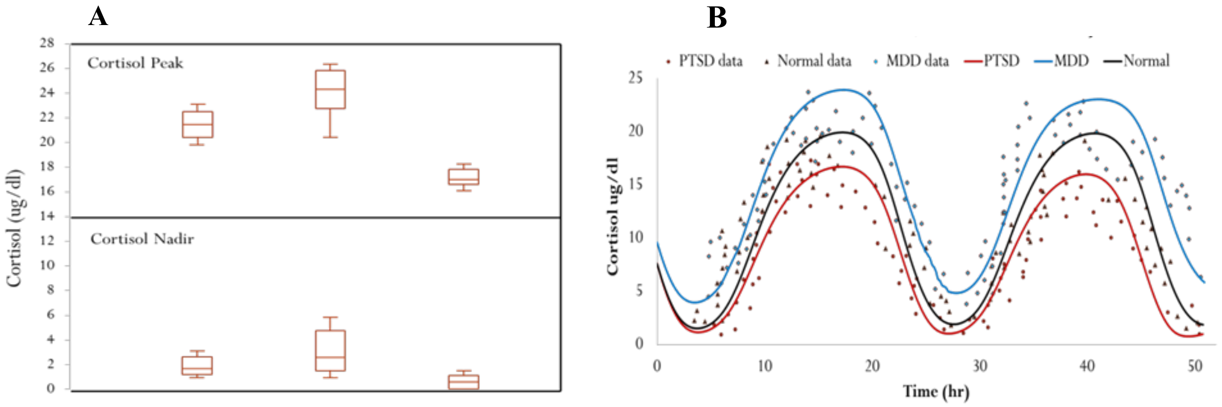


Figure 6.1 (A) Cortisol peak and nadir levels in healthy, MDD and PTSD subjects (B) Dynamic cortisol profiles of healthy, MDD and PTSD subjects. Reproduced from [69].

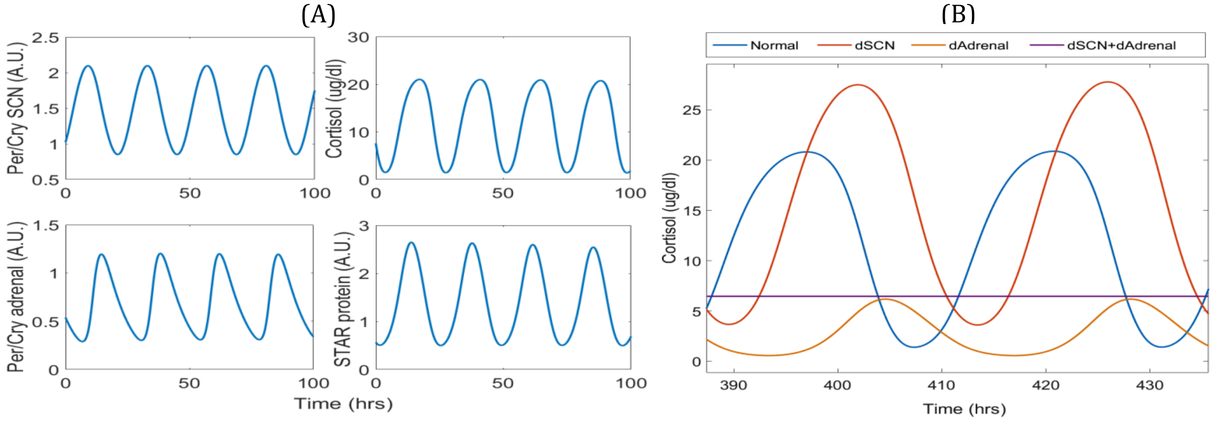


Figure 6.2 (A) Model dynamic for SCN, HPA and adrenal components. (B) Cortisol response under mutant scenarios.

6.2 Parametric Sensitivity Analysis

A local sensitivity analysis of the 235 model parameters was performed using MATLAB to assess the most sensitive parameters towards cortisol profiles. We analyzed the sensitivity of cortisol peak, nadir and period to obtain insights into MDD and PTSD specific cortisol phenotype. The sensitivity analysis reveals that the cortisol nadir level is highly sensitive, as compared to its peak and period. The parameters for kinetics of the Cry protein in SCN clock, steroidogenesis in HPA and the auxiliary loop in adrenal clock module showed highest sensitivity to cortisol profiles (**Figure 6.3**).

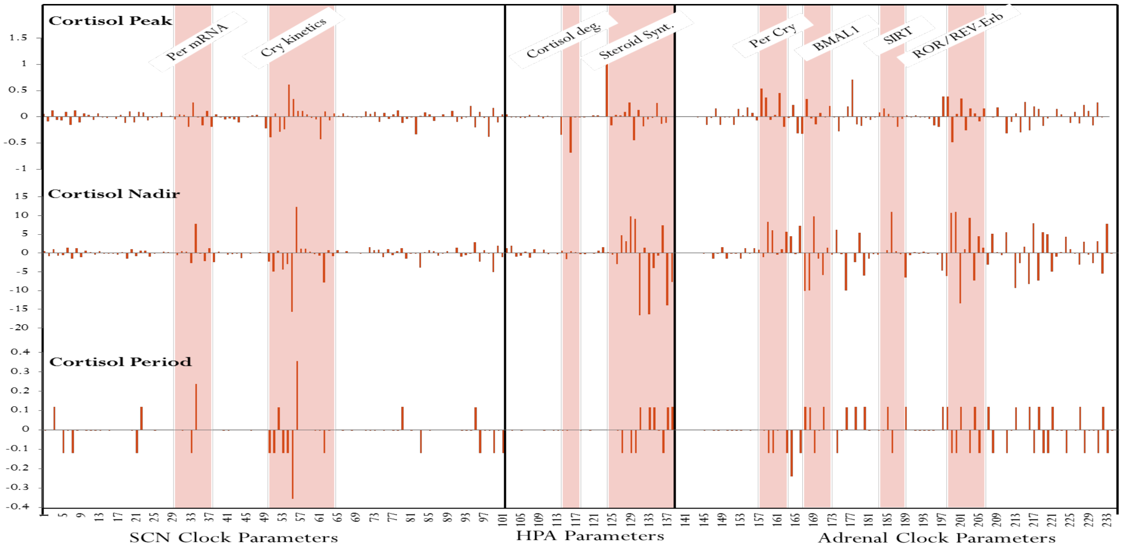


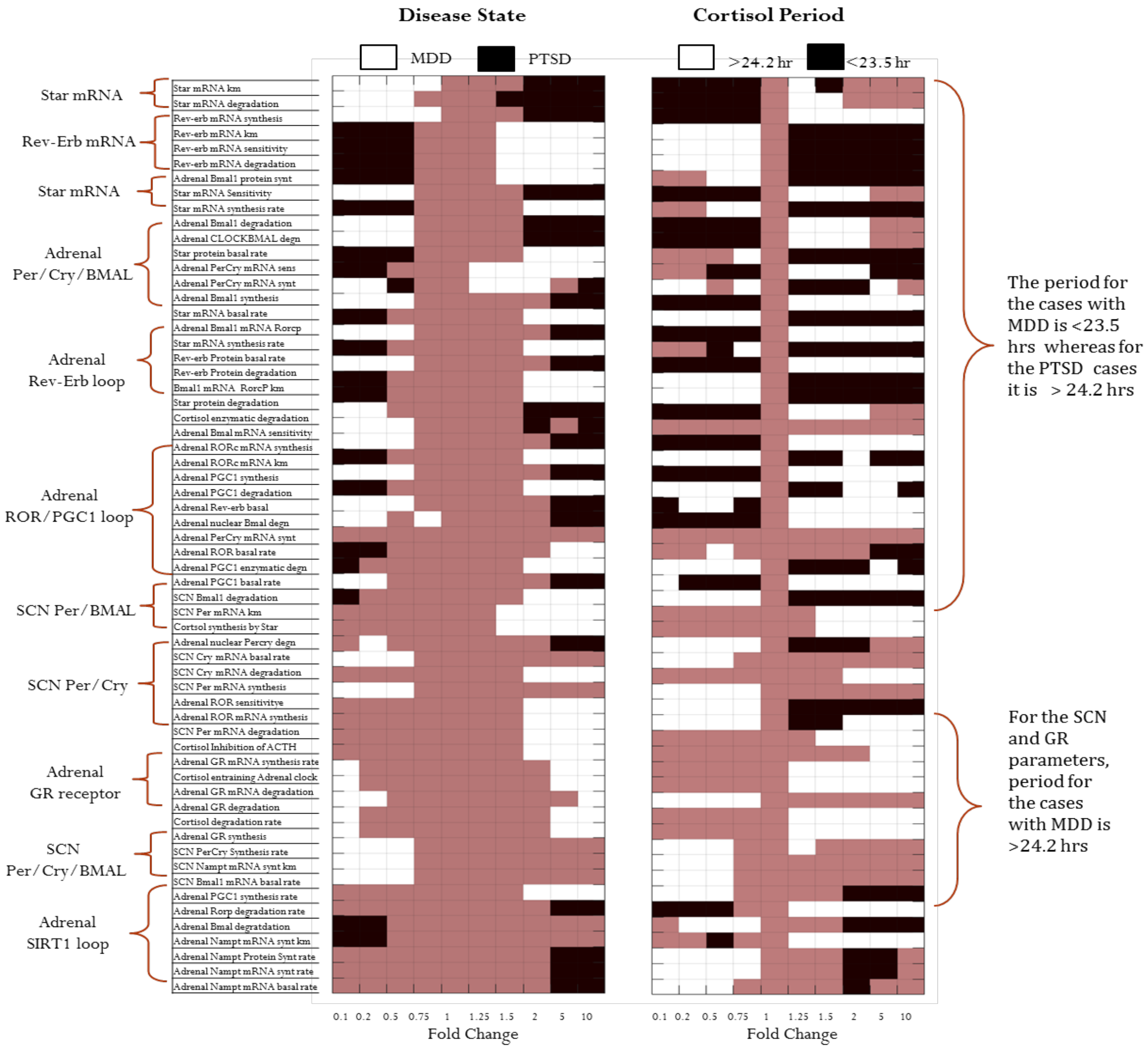
Figure 6.3 Sensitivity indices for model parameters with respect to cortisol peak, nadir and period.

6.3 Perturbation Analysis

A fold change perturbation analysis was performed for all the model parameters ranging from 0.1-10 fold. The mean of peak and nadir levels of cortisol oscillations were used to categorize MDD, PTSD or normal cortisol response. PTSD specific profiles were observed only for the parameter perturbations in the HPA axis and adrenal clock and not for the SCN clock. A higher perturbation was required to induce PTSD specific cortisol response as compared to MDD. While the period of MDD profiles due to adrenal perturbation was <23.5, the period for MDD profiles due to SCN perturbation was >24.2 hrs. The system elicits robust cortisol response over 25% of perturbations in the overall parameter space with only 7 parameters sensitive to stress in this range (**Figure 6.4**).

6.4 Insights into the etiology of the disease

The model analysis revealed the distinction in circadian pattern of plasma cortisol in MDD and PTSD with increased cortisol period for PTSD and decreased period in MDD. The model shows the decrease in circadian amplitude for MDD, which is in line with published data [74-75]. Cortisol nadir levels are highly sensitive as compared to its peak levels to the model parameters. The parameters with higher sensitivity can induce both MDD and PTSD specific cortisol profiles on perturbation in either direction. The processes of steroidogenesis and the auxiliary feedback loop in the adrenal clock are most likely to affect cortisol rhythm in neurological disorders. The cortisol profile is sensitive to depressive stress but relatively robust towards post-traumatic stress. The model suggests that the disruption in adrenal response is an essential etiology for PTSD. The model also shows that cortisol-mediated serotonergic drive is decreased in PTSD, implying the importance of serotonin supplementation in PTSD.



The period for the cases with MDD is <23.5 hrs whereas for the PTSD cases it is > 24.2 hrs

For the SCN and GR parameters, period for the cases with MDD is >24.2 hrs

Figure 6.4 Perturbation analysis showing cortisol response for sensitive parameters and the comparison of disease status with the cortisol period.

6.5 Statistical analysis of Multi-omics data from PTSD consortium

We performed statistical analysis on the data for metabolomics, clinical variables, neuro-endocrinology, physiology and proteomics data from the 82/82 cohort. To ascertain the stable features in the 82/82 datasets we performed bootstrapping on the data with 1000-fold random sampling of 50 of 82 subjects for control and PTSD groups. The randomly sampled 50/50 data sets were used for statistical analysis and the process was repeated for 1000 iterations. Several two-sample statistical tests, including student's t-test for significance in difference of means and Wilcox test for significance in difference of medians were performed to identify the key features that are different between PTSD and control subjects. The stable features were identified using the following filtration criteria: p<0.01 for difference in mean *or* median along with a mean or median fold change of at least 25% in at least 50% of the bootstrap iterations. Several features showed differences in the variances with significant mean differences after removal of the outliers. Such scenarios may not be captured using just statistical tests for mean or median. We accounted for statistically significant difference in the properties of distributions along with the significant differences in the fold change information to filter the features. Table 6.1 reports the features and the mean of the mean and median fold change (PTSD/Control) across 1000 random iterations.

Table 6.1 Summary of stable features with p<0.01 and a mean or median fold-change of at least 25% (lower or higher) in at least 50% of the 1000 bootstrapping iterations.

Omics variables	Mean fold change	Median fold change
Immune and Inflammation		
Cytokine IL-6	4.57	1.40
Cytokine TNF-alpha	1.24	1.13
Cytokine IL-12	0.28	1.03
Cytokine IFN-g	1.18	1.94
CD133+KDR+CD14+ monocyte	1.58	1.37
CD16nCD56p (% NK Cells)	0.79	0.85
CD16pCD56n (% NK Cells)	1.41	1.44
Clinical Lab		
Aspartate transaminase	1.26	1.18
Alanine Aminotransferase	1.26	1.00
Gamma-glutamyltransferase	1.14	1.36
C reactive protein	1.87	1.45
Neuroendocrine		
Cordif (cor1-cor2)	1.43	1.38
Urine Norepinephrine	1.18	1.11
5b-tetrahydrocortisol	0.83	0.77
Physiological		
pulse	1.12	1.13
Proteomics		
APOC4.AWF	1.42	1.55
C4BPB.LIQ	1.40	1.86
CRP.GYS	1.96	1.68
PRG4.DQY	1.29	1.22
Metabolomics		
Insulin	1.54	1.43
BDNF	1.18	1.17
m12propanediol	1.67	1.06
m13dipalmitoylglycerol	1.44	1.25
m5oxoproline	1.10	1.10
m7alphahydroxycholesterol	1.66	1.27
ADSGEGDFXAEGGGVR	3.59	1.90
cotinine	4.27	1.00
dihomolinoleate202n6	0.79	0.79
docosahexaenoateDHA226n3	0.78	0.82
DSGEGDFXAEGGGVR	6.08	2.66
eicosenoate201n9or11	0.75	0.85
hypoxanthine	1.34	1.31
lactate	1.29	1.32
Nacetylornithine	0.72	0.74
phenyllactatePLA	1.20	1.17
pyruvate	1.27	1.29

6.6 Analysis of Metabolomics data in PTSD: The source of lactate/pyruvate/FFA in PTSD

The analysis of the metabolomics data shows that lactate, pyruvate, citrate, and the urea cycle components are significantly higher in PTSD subjects. The analysis of the uptake and release rates for these key metabolites by 10 different tissue types in the human body reveals that muscle, liver and adipose tissues are most likely to render higher levels of these metabolites in blood (**Tables 6.2-6.3**). The statistical analysis reveals a net catabolic state in these tissues leading to such a metabolic phenotype. The higher levels of adrenaline/catacholamine can induce a catabolic state in the tissue leading to increased pyruvate/lactate, fatty acids and urea cycle. The analysis reveals that the muscle-liver-adipose axis can be affected due to higher adrenal/cortisol ratio.

Table 6.2 The physiological rates of metabolite transport for different tissues as reported in [76].

	For 70 kg man ~22 BMI			Metabolite flux (mmol/min)						
Tissue	Tissue Wt.	Blood flow	Vol. flow	Gluc	Pyr	Lact	FFA	Ala	TG	GLR
	kg	lit/min	lit/min/kg							
Brain	1.45	0.75	0.52	0.38	0	0	0	0	0	0
Liver	1.5	1.5	1.00	-0.731	0	0.27	0.21	0.32	-0.029	0.14
Muscle	20	0.9	0.05	0.038	0.005	-0.11	0.046	-0.04	0.003	-0.003
Adipose	11	0.36	0.03	0.04	0	-0.06	-0.211	0	0.02	-0.097
Heart	0.25	0.25	1.00	0.04	0	0.04	0.035	0	0	0
GI Track	2	1.1	0.55	0.076	0	0	-0.12	0	0.006	-0.04
Kidney	0.25	0.55	2.20	-0.06	0	0	0	-0.28	0	0
Lungs	0.7	0.25	0.36		0	0	0	0	0	0
Blood/RBC	2				-0.005	-0.14	0	0	0	0
Other/Bones	28.85	0.09								

Table 6.3 The net metabolite flux and metabolic pathways active in each tissue. Yellow highlighted fields represent the major contributing tissues to the net flux for glucose, pyruvate, lactate, fatty acids and active metabolic pathways.

Glysis: Glycolysis; Glnsis: Gluconeogenesis; Glyc.met: Glycogen metabolism; FFA synth: Fatty acid synthesis; Beta Oxdn: beta-oxidation; TG met: Triglyceride metabolism; Chol met: Cholesterol metabolism; Prot met: Protein metabolism; Insulin Depdt: Insulin dependent regulation.

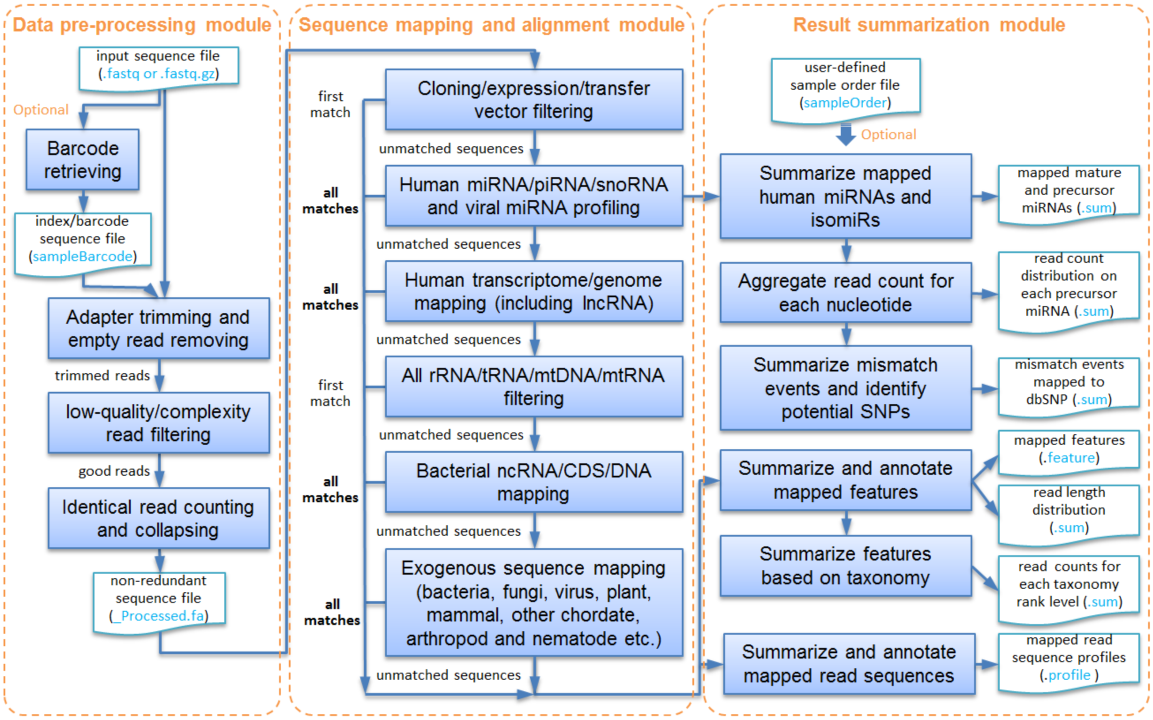
	Net flux (mmol/l/kg)				Metabolic Pathways										
Tissue	Glucose	Pyr	Lact.	FFA	Glysis	Glnsis	Glyc met	TCA cycle	Urea cycle	FFA synth	Beta oxdn	TG met	Chol met	Prot met	Insulin Depdt
Brain	0.735	0.000	0.000	0.000	🟨		🟨	🟨							
Liver	-0731	0.000	0.270	0.210	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Muscle	0.844	0.111	-2.489	1.022	🟨		🟨	🟨	🟨	🟨	🟨	🟨		🟨	🟨
Adipose	1.222	0.000	-1.711	-6.447	🟨					🟨	🟨	🟨			🟨
Heart	0.040	0.000	0.040	0.035	🟨		🟨	🟨		🟨	🟨	🟨	🟨		
GI Track	0.138	0.000	0.000	-0.218	🟨					🟨		🟨			
Kidney	-0.027	0.000	0.000	0.000	🟨	🟨	🟨		🟨					🟨	
Lungs	0.000	0.000	0.000	0.000	🟨										
Blood/RBC					🟨			🟨							
Other/Bones					🟨										

7. Development of data analysis pipelines for large molecular datasets

7.1 Development of Small RNA sequencing data analysis pipeline (sRNAnalyzer)

To achieve better coverage of the transcriptome, we improved the RNA analysis pipeline described in our earlier publication [77]. The new pipeline comprises three functional modules (**Figure 7.1**): data pre-processing, sequence mapping (alignment), and result summarization. Bowtie and Bowtie 2 aligners are used to handle both small and large RNA sequence mapping. The pipeline also includes both endogenous and exogenous RNA mapping steps. A local probabilistic model is used to assign reads to the most-likely sequence identity. Because of its modular design, the pipeline allows rapid modifications of each module without affecting the overall pipeline operation, for example adding a reference database or changing the order of databases used in read sequence alignment.

On selecting features to generate the biomarker panel, we applied support vector machine (SVM) with recursive feature elimination (SVM-RFE) algorithm [78] to find an optimal subset of features to classify patient and control groups. The SVM-RFE was applied to select peptides (proteins) and miRNAs to separate the PTSD- and PTSD+ groups. The 5-fold cross-validation was repeated 100 times to identify optimized features and to obtain an unbiased estimation of classification accuracy. The importance of each feature in the classification was determined based on the selection frequency from 5-fold cross-validations. The features were sorted in order of their frequencies. By increasing the number of features, SVM models were constructed and the average classification accuracies were computed. The optimal feature set was then determined at the highest average classification accuracy of the test set. For integrative analysis of SRM and miRNA-seq data, we scaled the data into z-scores separately and then concatenated the two datasets.



Data pre-processing module
input sequence file (.fastq or .fastq.gz)
Optional
Barcode retrieving
index/barcode sequence file (sampleBarcode)
Adapter trimming and empty read removing
trimmed reads
low-quality/complexity read filtering
good reads
Identical read counting and collapsing
non-redundant sequence file (_Processed.fa)
Sequence mapping and alignment module
first match
Cloning/expression/transfer vector filtering
unmatched sequences
all matches
Human miRNA/piRNA/snoRNA and viral miRNA profiling
unmatched sequences
all matches
Human transcriptome/genome mapping (including lncRNA)
unmatched sequences
first match
All rRNA/tRNA/mtDNA/mtRNA filtering
unmatched sequences
all matches
Bacterial ncRNA/CDS/DNA mapping
unmatched sequences
all matches
Exogenous sequence mapping (bacteria, fungi, virus, plant, mammal, other chordate, arthropod and nematode etc.)
unmatched sequences
Result summarization module
user-defined sample order file (sampleOrder)
Optional
Summarize mapped human miRNAs and isomiRs
mapped mature and precursor miRNAs (.sum)
Aggregate read count for each nucleotide
read count distribution on each precursor miRNA (.sum)
Summarize mismatch events and identify potential SNPs
mismatch events mapped to dbSNP (.sum)
Summarize and annotate mapped features
mapped features (.feature)
read length distribution (.sum)
Summarize features based on taxonomy
read counts for each taxonomy rank level (.sum)
Summarize and annotate mapped read sequences
mapped read sequence profiles (.profile)

Figure 7.1 Main framework of sRNAnalyzer. The pipeline can be divided into three functional modules which are separated by doted lines. The data format for each process is indicated in blue characters.

7.2 Automation of imputation and meta-analysis pipeline for GWAS data

With the ever increasing availability of densely genotyped reference genomes, it has become possible to impute large set of genetic variants from carefully chosen tagger-SNPs available on commercial arrays (**Figure 7.2**). However, this imputation process, aside from requiring huge computational resources, involves careful execution of several steps. We have prepared a set of bash script files to automate the imputation and meta-analysis process of genome studies (**Figure 7.3**). This will enable us to perform the GWAS analysis with a more dense coverage data, which may lead to improvement in power of association statistics, and help identify the exact location of the causative variants. Also, some widely studied SNPs in PTSD are not genotyped on Illumina's HumanOmniExpress BeadChip, and need to be imputed.

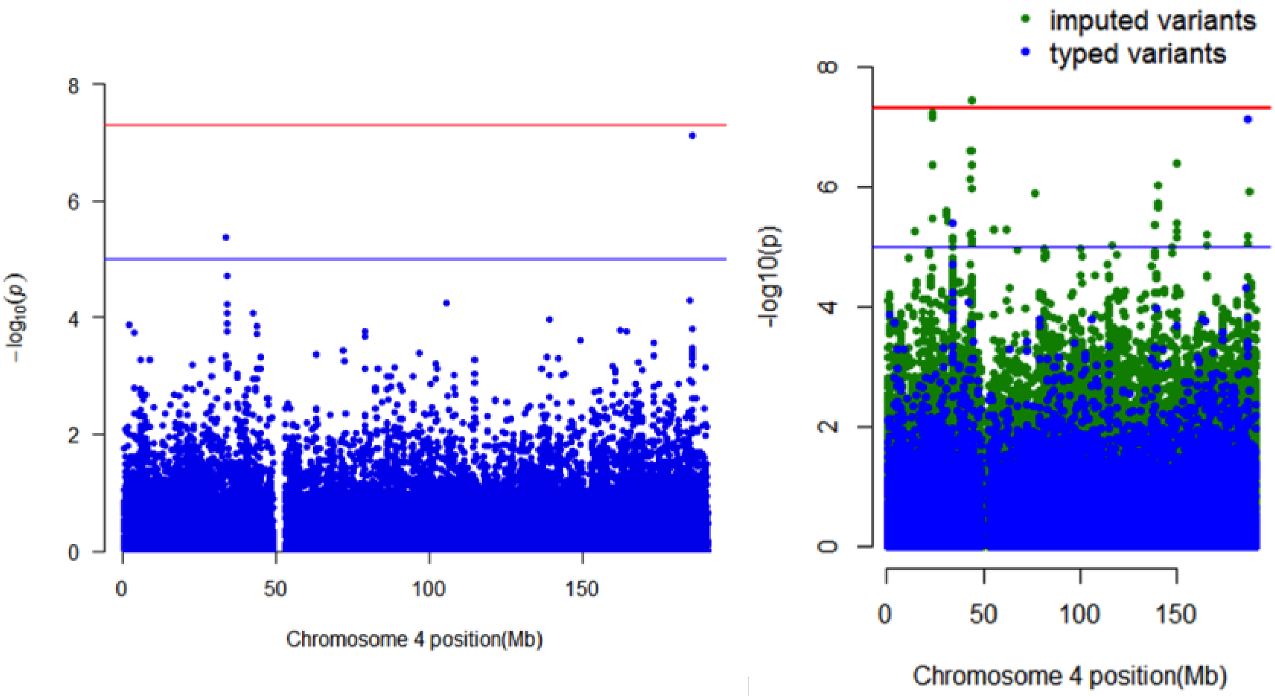


Figure 7.2 GWAS improvement from imputation in 147 subject dataset (shown only for Chromosome 4). Left: original GWAS manhattan plot, showing significance p-values ordered by chromosomal position. Right: improved statistical significance of genetic variants due to improved coverage by imputation.

Our imputation pipeline is summarized on the workflow diagram shown in Figure 7.3. First, strand orientation of genotyped data is checked and corrected with PLINK. Then, the data is split into individual chromosomes. As pre-phasing improves imputation accuracy and speed, the study data is pre-phased with SHAPEIT using genetic map data for build 37. Imputation is done for a window of 5Mb at a time with IMPUTE2 using phased reference panel from 1000 Genome Project phase 3 dataset. Then the imputed data is reassembled with GTOOL.

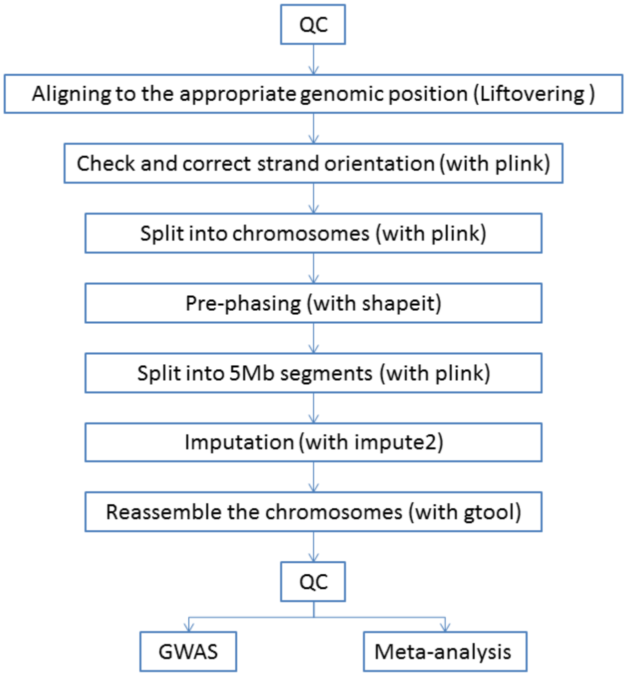

QC
Aligning to the appropriate genomic position (Liftovering)
Check and correct strand orientation (with plink)
Split into chromosomes (with plink)
Pre-phasing (with shapeit)
Split into 5Mb segments (with plink)
Imputation (with impute2)
Reassemble the chromosomes (with gtool)
QC
GWAS
Meta-analysis

Figure 7.3 Overview of automatic imputation pipeline.

References

1. T. M. Keane, B. P. Marx and D. M. Sloan, *Post-Traumatic Stress Disorder*, Springer, 2009, pp. 1–19
2. K. H. Seal, T. J. Metzler, K. S. Gima, D. Bertenthal, S. Maguen and C. R. Marmar, *Am. J. Public Health*, 2009, **99**, 1651–1658
3. Thakur, G.S., Daigle Jr, B.J., Dean, K.R., Zhang, Y., Rodriguez-Fernandez, M., Hammamieh, R., Yang, R., Jett, M., Palma, J., Petzold, L.R. and Doyle III, F.J., 2015. Systems biology approach to understanding post-traumatic stress disorder. *Molecular BioSystems*, *11*(4), pp.980-993.
4. Catalona, W.J., Smith, D.S., Ratliff, T.L. and Basler, J.W., 1993. Detection of organ-confined prostate cancer is increased through prostate-specific antigen—based screening. *Jama*, *270*(8), pp.948-954.
5. Katus, H.A., Remppis, A., Neumann, F.J., Scheffold, T., Diederich, K.W., Vinar, G., Noe, A., Matern, G. and Kuebler, W., 1991. Diagnostic efficiency of troponin T measurements in acute myocardial infarction. *Circulation*, *83*(3), pp.902-912.
6. Chen, X., Ba, Y., Ma, L., Cai, X., Yin, Y., Wang, K., Guo, J., Zhang, Y., Chen, J., Guo, X. and Li, Q., 2008. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell research*, *18*(10), pp.997-1006.
7. Mitchell, P.S., Parkin, R.K., Kroh, E.M., Fritz, B.R., Wyman, S.K., Pogosova-Agadjanyan, E.L., Peterson, A., Noteboom, J., O'Briant, K.C., Allen, A. and Lin, D.W., 2008. Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, *105*(30), pp.10513-10518.
8. Wang, K., Zhang, S., Marzolf, B., Troisch, P., Brightman, A., Hu, Z., Hood, L.E. and Galas, D.J., 2009. Circulating microRNAs, potential biomarkers for drug-induced liver injury. *Proceedings of the National Academy of Sciences*, *106*(11), pp.4402-4407.
9. Wang, G.K., Zhu, J.Q., Zhang, J.T., Li, Q., Li, Y., He, J., Qin, Y.W. and Jing, Q., 2010. Circulating microRNA: a novel potential biomarker for early diagnosis of acute myocardial infarction in humans. *European heart journal*, *31*(6), pp.659-666.
10. Denzer, K., Kleijmeer, M.J., Heijnen, H.F., Stoorvogel, W. and Geuze, H.J., 2000. Exosome: from internal vesicle of the multivesicular body to intercellular signaling device. *Journal of cell science*, *113*(19), pp.3365-3374.
11. Andaloussi, S.E., Mäger, I., Breakefield, X.O. and Wood, M.J., 2013. Extracellular vesicles: biology and emerging therapeutic opportunities. *Nature reviews Drug discovery*, *12*(5), pp.347-357.
12. Yoon, Y.J. and Gho, Y.S., 2014. Extracellular vesicles as emerging intercellular communicasomes. *BMB reports*, *47*(10), pp.531-539.
13. Kanada, M., Bachmann, M.H., Hardy, J.W., Frimannson, D.O., Bronsart, L., Wang, A., Sylvester, M.D., Schmidt, T.L., Kaspar, R.L., Butte, M.J. and Matin, A.C., 2015. Differential fates of biomolecules delivered to target cells via extracellular vesicles. *Proceedings of the National Academy of Sciences*, *112*(12), pp.E1433-E1442.
14. Koenen, K.C., Lyons, M.J., Goldberg, J., Simpson, J., Williams, W.M., Toomey, R., Eisen, S.A., True, W.R., Cloitr, M., Wolfe, J. and Tsuang, M.T., 2003. A high risk twin study of combat-related PTSD comorbidity. *Twin Research*, *6*(03), pp.218-226.
15. Almli, L.M., Stevens, J.S., Smith, A.K., Kilaru, V., Meng, Q., Flory, J., Abu-Amara, D., Hammamieh, R., Yang, R., Mercer, K.B. and Binder, E.B., 2015. A genome-wide identified risk variant for PTSD is a methylation quantitative trait locus and confers decreased cortical activation to fearful faces. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *168*(5), pp.327-336.
16. Logue, M.W., Baldwin, C., Guffanti, G., Melista, E., Wolf, E.J., Reardon, A.F., Uddin, M., Wildman, D., Galea, S., Koenen, K.C. and Miller, M.W., 2013. A genome-wide association study of post-traumatic stress disorder identifies the retinoid-related orphan receptor alpha (RORA) gene as a significant risk locus. *Molecular psychiatry*, *18*(8), pp.937-942.
17. Hunter, R.G., McCarthy, K.J., Milne, T.A., Pfaff, D.W. and McEwen, B.S., 2009. Regulation of hippocampal H3 histone methylation by acute and chronic stress. *Proceedings of the National Academy of Sciences*, *106*(49), pp.20912-20917.
18. Jergović, M., Tomičević, M., Vidović, A., Bendelja, K., Savić, A., Vojvoda, V., Rac, D., Lovrić-Čavar, D., Rabatić, S., Jovanovic, T. and Sabioncello, A., 2014. Telomere shortening and immune activity in war veterans with posttraumatic stress disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *54*, pp.275-283.

19. Lohr, J.B., Palmer, B.W., Eidt, C.A., Aailaboyina, S., Mausbach, B.T., Wolkowitz, O.M., Thorp, S.R. and Jeste, D.V., 2015. Is post-traumatic stress disorder associated with premature senescence? A review of the literature. *The American Journal of Geriatric Psychiatry*, *23*(7), pp.709-725.
20. Čeprnja, M., Đerek, L., Unić, A., Blažev, M., Fistonić, M., Kozarić-Kovačić, D., Franić, M. and Romić, Ž., 2011. Oxidative stress markers in patients with post-traumatic stress disorder. *Collegium antropologicum*, *35*(4), pp.1155-1160.
21. Bersani, F.S., Morley, C., Lindqvist, D., Epel, E.S., Picard, M., Yehuda, R., Flory, J., Bierer, L.M., Makotkine, I., Abu-Amara, D. and Coy, M., 2016. Mitochondrial DNA copy number is reduced in male combat veterans with PTSD. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *64*, pp.10-17.
22. Almli, L.M., Stevens, J.S., Smith, A.K., Kilaru, V., Meng, Q., Flory, J., Abu-Amara, D., Hammamieh, R., Yang, R., Mercer, K.B., Binder, E.B., Bradley, B., Hamilton, S., Jett, M., Yehuda, R., Marmar, C.R., and Ressler, K.J., 2015. A genome-wide identified risk variant for PTSD is a methylation quantitative trait locus and confers decreased cortical activation to fearful faces. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *168*(5), pp.327-336.
23. Almli, L.M., Fani, N., Smith, A.K. and Ressler, K.J., 2014. Genetic approaches to understanding post-traumatic stress disorder. *International Journal of Neuropsychopharmacology*, *17*(2), pp.355-370.
24. Voisey, J., Young, R.M., Lawford, B.R. and Morris, C.P., 2014. Progress towards understanding the genetics of posttraumatic stress disorder. *Journal of Anxiety Disorders*, *28*(8), pp.873-883.
25. Nievergelt, C.M., Maihofer, A.X., Mustapic, M., Yurgil, K.A., Schork, N.J., Miller, M.W., Logue, M.W., Geyer, M.A., Risbrough, V.B., O'Connor, D.T. and Baker, D.G., 2015. Genomic predictors of combat stress vulnerability and resilience in US Marines: A genome-wide association study across multiple ancestries implicates PRTFDC1 as a potential PTSD gene. *Psychoneuroendocrinology*, *51*, pp.459-471.
26. Guffanti, G., Galea, S., Yan, L., Roberts, A.L., Solovieff, N., Aiello, A.E., Smoller, J.W., De Vivo, I., Ranu, H., Uddin, M. and Wildman, D.E., 2013. Genome-wide association study implicates a novel RNA gene, the lincRNA AC068718. 1, as a risk factor for post-traumatic stress disorder in women. *Psychoneuroendocrinology*, *38*(12), pp.3029-3038.
27. Logue, M.W., Baldwin, C., Guffanti, G., Melista, E., Wolf, E.J., Reardon, A.F., Uddin, M., Wildman, D., Galea, S., Koenen, K.C. and Miller, M.W., 2013. A genome-wide association study of post-traumatic stress disorder identifies the retinoid-related orphan receptor alpha (RORA) gene as a significant risk locus. *Molecular Psychiatry*, *18*(8), pp.937-942.
28. Xie, P., Kranzler, H.R., Yang, C., Zhao, H., Farrer, L.A. and Gelernter, J., 2013. Genome-wide association study identifies new susceptibility loci for posttraumatic stress disorder. *Biological Psychiatry*, *74*(9), pp.656-663.
29. Stein, M.B., Chen, C.Y., Ursano, R.J., Cai, T., Gelernter, J., Heeringa, S.G., Jain, S., Jensen, K.P., Maihofer, A.X., Mitchell, C. and Nievergelt, C.M., 2016. Genome-wide Association Studies of Posttraumatic Stress Disorder in 2 Cohorts of US Army Soldiers. *JAMA Psychiatry*.
30. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), pp.559-575.
31. Zhang, L., Benedek, D.M., Fullerton, C.S., Forsten, R.D., Naifeh, J.A., Li, X.X., Hu, X.Z., Li, H., Jia, M., Xing, G.Q. and Benevides, K.N., 2014. PTSD risk is associated with BDNF Val66Met and BDNF overexpression. *Molecular Psychiatry*, *19*(1), pp.8-10.
32. Schell, T.L. and Marshall, G.N., 2008. Survey of individuals previously deployed for OEF/OIF. *Invisible wounds of war: Psychological and cognitive injuries, their consequences, and services to assist recovery*, pp.87-115.
33. Kulka, R.A., Schlenger, W.E., Fairbank, J.A., Hough, R.L., Jordan, B.K., Marmar, C.R. and Weiss, D.S., 1990. *Trauma and the Vietnam war generation: Report of findings from the National Vietnam Veterans Readjustment Study*. Brunner/Mazel.
34. Pole, N., Best, S.R., Metzler, T. and Marmar, C.R., 2005. Why are Hispanics at greater risk for PTSD?. *Cultural Diversity and Ethnic Minority Psychology*, *11*(2), p.144.
35. Galea, S., Ahern, J., Resnick, H., Kilpatrick, D., Bucuvalas, M., Gold, J. and Vlahov, D., 2002. Psychological sequelae of the September 11 terrorist attacks in New York City. *New England Journal of Medicine*, *346*(13), pp.982-987.

36. Perilla, J.L., Norris, F.H. and Lavizzo, E.A., 2002. Ethnicity, culture, and disaster response: Identifying and explaining ethnic differences in PTSD six months after Hurricane Andrew. *Journal of Social and Clinical Psychology*, *21*(1), p.20.
37. Shabalin, A.A., 2012. Matrix eQTL: ultrafast eQTL analysis via large matrix operations. *Bioinformatics*, *28*(10), pp.1353-1358.
38. Langfelder, P. and Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, *9*(1), p.559.
39. Huang, D.W., Sherman, B.T. and Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, *4*(1), pp.44-57.
40. Huang, D.W., Sherman, B.T. and Lempicki, R.A., 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, *37*(1), pp.1-13.
41. Wang, J., Duncan, D., Shi, Z. and Zhang, B., 2013. WEB-based gene set analysis toolkit (WebGestalt): update 2013. *Nucleic acids research*, *41*(W1), pp.W77-W83.
42. Chen, Y., Li, X., Kobayashi, I., Tsao, D. and Mellman, T.A., 2016. Expression and methylation in posttraumatic stress disorder and resilience; evidence of a role for odorant receptors. *Psychiatry Research*, *245*, pp.36-44.
43. Chen, Y.C., Pal, N.R. and Chung, I.F., 2012. An integrated mechanism for feature selection and fuzzy rule extraction for classification. *IEEE Transactions on Fuzzy Systems*, *20*(4), pp.683-698.
44. Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., Shen, E.H., Ng, L., Miller, J.A., Van De Lagemaat, L.N., Smith, K.A., Ebbert, A., Riley, Z.L. and Abajian, C., 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, *489*(7416), pp.391-399.
45. "Allen Brain Atlas: Human Brain." Allen Institute for Brain Science, 2015. Web. http://human.brain-map.org/
46. Miller, J.A., Ding, S.L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K. and Arnold, J.M., 2014. Transcriptional landscape of the prenatal human brain. *Nature*, *508*(7495), pp.199-206.
47. "BrainSpan: Atlas of the Developing Human Brain." Allen Institute for Brain Science, 2015. Web. http://brainspan.org/
48. Yang, R., Daigle, B.J., Petzold, L.R. and Doyle, F.J., 2012. Core module biomarker identification with network exploration for breast cancer metastasis. *BMC bioinformatics*, *13*(1), p.1.
49. Thakur, G.S., Daigle, B.J., Qian, M., Dean, K.R., Zhang, Y., Yang, R., Kim, T.K., Wu, X., Li, M., Lee, I., Petzold, L.R., and Doyle III, F.J., A Multi-Metric Evaluation of Stratified Random Sampling for Classification: A Case Study. *IEEE Life Science Letters, in press.*
50. Surinova, S., Hüttenhain, R., Chang, C.Y., Espona, L., Vitek, O. and Aebersold, R., 2013. Automated selected reaction monitoring data analysis workflow for large-scale targeted proteomic studies. *Nature protocols*, *8*(8), pp.1602-1619.
51. Tian, G., Yin, X., Luo, H., Xu, X., Bolund, L. and Zhang, X., 2010. Sequencing bias: comparison of different protocols of microRNA library construction. *BMC biotechnology*, *10*(1), p.1.
52. Backes, C., Sedaghat-Hamedani, F., Frese, K., Hart, M., Ludwig, N., Meder, B., Meese, E. and Keller, A., 2016. Bias in high-throughput analysis of miRNAs and implications for biomarker studies. *Analytical chemistry*, *88*(4), pp.2088-2095.
53. Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J.T., Nusbaum, J.D., Morozov, P., Ludwig, J. and Ojo, T., 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *Rna*, *17*(9), pp.1697-1712.
54. Fuchs, R.T., Sun, Z., Zhuang, F. and Robb, G.B., 2015. Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. *PloS one*, *10*(5), p.e0126049.
55. Feusner, J., Ritchie, T., Lawford, B., Young, R.M., Kann, B. and Noble, E.P., 2001. GABA A receptor β3 subunit gene and psychiatric morbidity in a post-traumatic stress disorder population. *Psychiatry research*, *104*(2), pp.109-117.
56. Li, Y., Han, F. and Shi, Y., 2013. Increased neuronal apoptosis in medial prefrontal cortex is accompanied with changes of Bcl-2 and Bax in a rat model of post-traumatic stress disorder. *Journal of Molecular Neuroscience*, *51*(1), pp.127-137.
57. Muhie, S., Gautam, A., Meyerhoff, J., Chakraborty, N., Hammamieh, R. and Jett, M., 2015. Brain transcriptome profiles in mouse model simulating features of post-traumatic stress disorder. *Molecular brain*, *8*(1), p.1.
58. Rothbaum, B.O., Kearns, M.C., Reiser, E., Davis, J.S., Kerley, K.A., Rothbaum, A.O., Mercer, K.B., Price, M., Houry, D. and Ressler, K.J., 2014. Early intervention following trauma may mitigate genetic risk for PTSD in civilians: a pilot prospective emergency department study. *The Journal of clinical psychiatry*, *75*(12), pp.1380-1387.

59. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), pp.389-422.
60. Netpath. http://www.netpath.org/.
61. Atlas of Cancer Genes. http://atlasgeneticsoncology.org/.
62. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S. and Kok, C.Y., 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, *43*(D1), pp.D805-D811.
63. Mosca, E., Alfieri, R., Merelli, I., Viti, F., Calabria, A. and Milanesi, L., 2010. A multilevel data integration resource for breast cancer study. *BMC systems biology*, *4*(1), p.76.
64. Kanehisa, M. and Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *28*(1), pp.27-30.
65. Yu, X., Zeng, T. and Li, G., 2015. Integrative enrichment analysis: a new computational method to detect dysregulated pathways in heterogeneous samples. *BMC genomics*, *16*(1), p.1.
66. Shi, L., Lei, X. and Zhang, A., 2011. Protein complex detection with semi-supervised learning in protein interaction networks. *Proteome Science*, *9*(1), p.1.
67. Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T. and Lee, D., 2008. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, *4*(11), p.e1000217.
68. Drier, Y., Sheffer, M. and Domany, E., 2013. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, *110*(16), pp.6388-6393.
69. Yehuda, R., Teicher, M.H., Trestman, R.L., Levengood, R.A. and Siever, L.J., 1996. Cortisol regulation in posttraumatic stress disorder and major depression: a chronobiological analysis. *Biological Psychiatry*, *40*(2), pp.79-88.
70. Sriram, K., Rodriguez-Fernandez, M. and Doyle III, F.J., 2012. Modeling cortisol dynamics in the neuro-endocrine axis distinguishes normal, depression, and post-traumatic stress disorder (PTSD) in humans. *PLoS Computational Biology*, *8*(2), p.e1002379.
71. Foteinou, P., and Doyle III, F.J., 2012. A Cell Autonomous Circadian-Enzymatic Model Elucidates the Effects of SIRT1 on Circadian Amplitude, *Society for Research on Biological Rhythms Meeting.*
72. Hosseinichimeh, N., Rahmandad, H. and Wittenborn, A.K., 2015. Modeling the hypothalamus–pituitary–adrenal axis: a review and extension. *Mathematical Biosciences*, *268*, pp.52-65.
73. Mavroudis, P.D., Corbett, S.A., Calvano, S.E. and Androulakis, I.P., 2014. Mathematical modeling of light-mediated HPA axis activity and downstream implications on the entrainment of peripheral clock genes. *Physiological Genomics*, *46*(20), pp.766-778.
74. Son, G.H., Chung, S., Choe, H.K., Kim, H.D., Baik, S.M., Lee, H., Lee, H.W., Choi, S., Sun, W., Kim, H. and Cho, S., 2008. Adrenal peripheral clock controls the autonomous circadian rhythm of glucocorticoid by causing rhythmic steroid production. *Proceedings of the National Academy of Sciences*, *105*(52), pp.20970-20975.
75. Li, J.Z., Bunney, B.G., Meng, F., Hagenauer, M.H., Walsh, D.M., Vawter, M.P., Evans, S.J., Choudary, P.V., Cartagena, P., Barchas, J.D. and Schatzberg, A.F., 2013. Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. *Proceedings of the National Academy of Sciences*, *110*(24), pp.9950-9955.
76. Logan, R.W., Edgar, N., Gillman, A.G., Hoffman, D., Zhu, X. and McClung, C.A., 2015. Chronic stress induces brain region-specific alterations of molecular rhythms that correlate with depression-like behavior in mice. *Biological Psychiatry*, *78*(4), pp.249-258.
77. Kim, J., Saidel, G.M. and Cabrera, M.E., 2007. Multi-scale computational model of fuel homeostasis during exercise: effect of hormonal control. *Annals of Biomedical Engineering*, *35*(1), pp.69-90.
78. Wang, K., Li, H., Yuan, Y., Etheridge, A., Zhou, Y., Huang, D., Wilmes, P. and Galas, D., 2012. The complex exogenous RNA spectra in human plasma: an interface with human gut biota?. *PloS one*, *7*(12), p.e51009.

Publications, Talks and Posters from Harvard, ISB and UCSB Teams

Publications

Sriram, K., Rodriguez-Fernandez, M. and Doyle III, F.J., 2012. A detailed modular analysis of heat-shock protein dynamics under acute and chronic stress and its implication in anxiety disorders. *PloS one*, *7*(8), p.e42958.

Sriram, K., Rodriguez-Fernandez, M. and Doyle III, F.J., 2012. Modeling cortisol dynamics in the neuro-endocrine axis distinguishes normal, depression, and post-traumatic stress disorder (PTSD) in humans. *PLoS Comput Biol*, *8*(2), p.e1002379.

Yang, R., Daigle, B.J., Petzold, L.R. and Doyle, F.J., 2012. Core module biomarker identification with network exploration for breast cancer metastasis. *BMC bioinformatics*, *13*(1), p.1.

Cho, J.H., Lin, A. and Wang, K., 2013. Kernel-based method for feature selection and disease diagnosis using transcriptomics data. *Systems Biomedicine*, *1*(4), pp.254-260.

Yang, R., Daigle Jr, B.J., Muhie, S.Y., Hammamieh, R., Jett, M., Petzold, L. and Doyle, F.J., 2013. Core modular blood and brain biomarkers in social defeat mouse model for post traumatic stress disorder. *BMC systems biology*, *7*(1), p.1.

Chevillet, J.R., Lee, I., Briggs, H.A., He, Y. and Wang, K., 2014. Issues and prospects of microRNA-based biomarkers in blood and other body fluids. *Molecules*, *19*(5), pp.6080-6105.

Cho, J.H., Lee, I., Hammamieh, R., Wang, K., Baxter, D., Scherler, K., Etheridge, A., Kulchenko, A., Gautam, A., Muhie, S. and Chakraborty, N., 2014. Molecular evidence of stress-induced acute heart injury in a mouse model simulating posttraumatic stress disorder. *Proceedings of the National Academy of Sciences*, *111*(8), pp.3188-3193.

Lausted, C., Lee, I., Zhou, Y., Qin, S., Sung, J., Price, N.D., Hood, L. and Wang, K., 2014. Systems approach to neurodegenerative disease biomarker discovery. *Annual review of pharmacology and toxicology*, *54*, pp.457-481.

Thakur, G.S., Daigle, B.J., Petzold, L.R. and Doyle, F.J., 2014. A Multivariate Ensemble Approach for Identification of Biomarkers: Application to Breast Cancer. *IFAC Proceedings Volumes*, *47*(3), pp.809-814.

Thakur, G.S., Daigle Jr, B.J., Dean, K.R., Zhang, Y., Rodriguez-Fernandez, M., Hammamieh, R., Yang, R., Jett, M., Palma, J., Petzold, L.R. and Doyle III, F.J., 2015. Systems biology approach to understanding post-traumatic stress disorder. *Molecular BioSystems*, *11*(4), pp.980-993.

Thakur, G.S., Daigle, B.J., Qian, M., Dean, K.R., Zhang, Y., Yang, R., KIm, T.K., Wu, X., Li, M., Lee, I. and Petzold, L.R., 2016. A Multi-Metric Evaluation of Stratified Random Sampling for Classification: A Case Study. *IEEE Life Sciences Letters*.

Submitted papers

- **Abolfazl Doostparast Torshizi** and Linda R. Petzold. Graph-based Semi-Supervised Learning with Genomic Data Integration Using Condition-Responsive Genes Applied to Phenotype Classification. BMC Bioinformatics, *under review*.

- **Rasha Hammamieh, Nabarun Chakraborty, Seid Muhie, Ruoting Yang, Aarti Gautam,** Raina Kumar, Bernie Daigle, Yuanyang Zhang, Duna Abu Amara, Stacy-Ann Miller, Seshmalini Srinivasan, Rachel Yehuda, Linda Petzold, Frank Doyle, Owen Wolkowitz, Sindy Mellon, Charles Marmar, and Marti Jett. Whole Genome DNA Methylation Status Associated with Clinical PTSD Measures of OIF/OEF Veterans. Biological Psychiatry, *submitted*.

Presentations:

Talks:

- **Ruoting Yang**, Bernie J Daigle Jr, Linda R Petzold, Francis J Doyle III. **Core module network construction for breast cancer metastasis.** Presented at 10th World Congress on Intelligent Control and Automation (WCICA). Beijing, China. July 6-8 2012

- **Gunjan S Thakur**, Bernie J Daigle Jr, Linda R Petzold, Francis J Doyle III. **A multivariate ensemble approach for identification of biomarkers: application to breast cancer.** Presented at 19th World Congress of the International Federation of Automatic Control (IFAC). Cape Town, South Africa. August 24-29 2014.

- **Kai Wang. Characterization of cell-free RNA in circulation** Presented at Molecular Tri-conference. San Francisco, CA. March 16, 2016.

- **Kai Wang. The need of standardization on measuring circulating RNA.** Presented at Molecular Tri-conference. San Francisco, March 16, 2016.

- **Yong Zhou**. **Organ-specific proteins in biomarker discovery.** Presented at Biomarker Summit. San Diego, CA. March 21-23, 2016.

- **Min Young Lee**. **Discovery of integrative biomarkers for Post-traumatic stress disorder with brain imaging, endocrine, and proteomic features.** Presented at the Cascadia Proteomics Symposium. Seattle, WA. July 11-12, 2016.

Posters:

- **Kelsey R Dean**, Francis J Doyle III.
Title: Incorporating cell mixture information into batch correction improves cell type-specific functional enrichment. Presented at 5th International Conferences on Foundations of Systems Biology (FOSBE). August 2015.

- **Gunjan S Thakur,** Bernie J Daigle Jr, Kelsey R Dean, Linda R Petzold, Francis J Doyle III.
Title: Metric Focused Feature selection for customized biomarker identification. Presented at 5th International Conferences on Foundations of Systems Biology (FOSBE). August 2015.

- **Yong Zhou,** Shizhen Qin, Li Gray, Xiaowei Yan, Inyoul Lee, Kai Wang, Mary Brunkow, Leroy Hood.
Title: Organ-specific proteins in biomarker discovery. Presented at Institute for Systems Biology's 15th International Symposium. Emerging Technologies. April 4-5, 2016.

- **Inyoul Lee,** Taek-Kyun Kim, Yong Zhou, Shizhen Qin, Ji-Hoon Cho, Kelsey Scherler, David Baxter, Li Gray, Xiaogang Wu, Minyoung Lee, Aarti Gautam, Rasha Hammamieh, Rachel Yehuda, Charles Marmar, Marti Jett, and Kai Wang, Leroy Hood.
Title: A systems Approach to blood biomarker discovery for posttraumatic disorder (PTSD). Presented at Institute for Systems Biology's 15th International Symposium. Emerging Technologies. April 4-5, 2016.

- **David Baxter,** Xiaogang Wu, Taek-Kyun Kim, I Kelsey Scherler, Alton Etheridge, Kai Wang, and Leroy Hood.
  Title: Improvement to small RNA sequencing library construction. Presented at Institute for Systems Biology's 15th International Symposium. Emerging Technologies. April 4-5, 2016.

- **Xiaogang Wu**, Taek-Kyun Kim, David Baxter, Kelsey Scherler, Inyoul Lee, Kathie Walters, Kai Wang, Leroy Hood
  Title: sRNAnalyzer - A flexible and customizable small RNA sequencing data analysis pipeline. Presented at Institute for Systems Biology's Annual International Symposium. Emerging Technologies. April 4-5, 2016. Presented at Institute for Systems Biology's 15th International Symposium. Emerging Technologies. April 4-5, 2016.

- **Taek-Kyun Kim**, Kai Wang, Xiaogang Wu, Daehee Hwang, Inyoul Lee, and Leroy Hood.
  Title: BASE: A tool to determine spatial distribution of genes in the mouse brain. Presented at Institute for Systems Biology's 15th International Symposium. Emerging Technologies. April 4-5, 2016.

- **Pramod R Somvanshi**, Panagiota Foteinou, and Francis J Doyle III.
  Title: Sensitivity of Circadian Cortisol Profiles in Neuro-endocrine Circuit: Analysis for Major Depressive Disorder and Post Traumatic Stress Disorder. Presented at Systems Biology of Human Disease. June 14-16, 2016

- **Kelsey R Dean** and Francis J Doyle III.
  Title: Identification and Characterization of Posttraumatic Stress Disorder Subtypes from Genome-Wide DNA Methylation Patterns. Presented at Systems Biology of Human Disease. June 14-16, 2016